社団法人 情報処理学会 研究報告
IPSJ SIG Technical Reports

2007－DBS－143 (33)
2007／7／3

キーワード連動広告でのキーワード発見手法の提案

岩切進悟† 下司義寛† 廣川佐千男††

† 九州大学システム情報科学府 〒819-0395 福岡市西区大字元岡 744
†† 九州大学情報基盤研究開発センター 〒812-8581 福岡市東区箱崎 6-10-1
E-mail: †{s-iwa,y-shimo}@i.kyushu-u.ac.jp, ††hirokawa@cc.kyushu-u.ac.jp

あらまし 検索キーワードのログを文書群とみなし, そこに現れる単語の共起情報についての概念グラフを求めることにより, 検索連動広告のための新しいキーワードを発見する手法を提案する. 検索エンジン, 求人サイト, ショッピングサイトのアクセスログの3つの検索ログにおける単語の共起率に基づき, 関連語抽出と, それらの関連性可視化により定性的評価を行った。 また, 検索エンジンでのログとサイト・アクセスログにおける単語の共起比率の差を利用しキーワード広告における機会損失評価法を提案する。

キーワード 検索連動広告 概念グラフ キーワード発見

Keyword Mining for Pay Per Click (PPC) using Query Logs

Shingo IWAKIRI†, Yoshihiro SHIMOJI†, and Sachio HIROKAWA††

† Graduate School of Information Science and Electrical Engineering, Kyushu University
†† Research Institute of Information Technology, Kyushu University
E-mail: †{s-iwa,y-shimo}@i.kyushu-u.ac.jp, ††hirokawa@cc.kyushu-u.ac.jp

Abstract This paper proposes a keyword discovery method based on the notion of the concept graph which extracts relationship between key words that appear in the same session of the query. Qualitative evaluations are shown using query logs of a general search engine, a job site and a shopping site. Another method, which uses the difference of the co-occurrence probability of two words, is proposed for quantitative evaluation of effectiveness of the proposed keywords.

Key words Paid Listing, Concept Graph, Keyword Discovery

1. まえがき

検索エンジンの利用者がどのようなクエリで検索を行っているかを理解することは, 検索エンジン事業者の最も大きな感心事である。商用検索エンジンが始まった当初から, 検索エンジンに蓄積される検索ログの分析は行われている。例えば, [3]では9種類の検索エンジンのクエリログについて, クエリの出現頻度, クエリ中の単語数, カテゴリごとの比較などの分析が行われている。[2]では, 検索の背景にある利用者の意図に基づき, クエリを誘導的, 情報収集的, トランザクション的の三種類に分類することを試みている。クエリ分類の研究は, KDD CUP2005 [5](注1) のテーマとしても取り上げられ, その後も引続き活発な研究がなされている [1], [4], [6]。しかし, これらの研究は, 商用検索エンジンが分野ごとの複数のバックエンド検索エンジンから構成され, 分類されるクエリを該当するバックエンド検索エンジンに分割処理することが目的である。

ところが, 検索連動広告が商用検索エンジンの中心的ビジネスモデルとなった現在, このようなクエリの一般的な分析だけでなく, より詳細な分析が必要となっている。検索連動広告とは, 検索エンジンの検索結果において検索クエリに対応した広告を表示する広告手法であり, Google アドワーズや Yahoo スポンサードサーチが検索連動広告の例である。広告出稿者は, 自社の商品やサービスに関連する適切なキーワードに対して広告の出稿を行い, そのキーワードと検索クエリが一致した場合に広告が表示される。これにより, 検索エンジン利用者を効率良く自サイトに誘導することができる。検索者が広告をクリックした時に広告出稿者が支払う金額を入札によって決め, その結果に基づいて広告の表示順位が決まる。つまり, キーワードをひとつの商品のように見立て, 広告としての価値が入札により決定される。図1の入札の例では「転職」というキーワードに対し, 最上位に表示させるためのコストとして1クリックあたり711円の値が付いている。検索結果の広告欄は, この入札結果順が反映される。

検索連動広告の出稿者はどのようなキーワードに対して広告

(注1) : http://www.acm.org/sigs/sigkdd/kdd2005/kddcup.html

図 1　入札価格一覧

を出稿するかが問題となる。SEM（検索エンジンマーケティング）を支援する企業では蓄積されたノウハウや人の力に大きく依存しながらキーワード発見のコンサルティングを行っている。個人サイトに対しても様々な工夫がなされている [9]。これらの手法を用い試行錯誤でキーワードを決められているが，いずれも手間と時間がかかる。検索連動広告のためのキーワード発見について，いくつかの企業が自動化を掲げているが詳細な技術的背景は明らかにされていない。[7], [10] では，クエリの分類ではなく，頻繁に検索されるクエリが具体的にどのようなものか分析している。[6] は他の利用者のクエリから抽出されるキーワードを検索拡張として利用者に提示することで検索の効率が向上できることを示している。しかし，著者等が知る限りでは，検索連動広告出稿者が広告を出稿すべきキーワードをどのように選択すべきかを考察する研究はない。

筆者等は，文書群に現れる単語の文書頻度に基づき単語間の上位下位関係を抽出し可視化する概念グラフを提案してきた [8]。本稿では，この概念グラフの理論を検索ログに適用することにより，検索連動広告において適切なキーワードを発見する方式を提案し，実データについて行った定性的評価について述べる。また，一般の検索エンジンにおけるクエリログと特定サイトへのアクセスログ中のクエリログにおける単語間の共起比率の差を利用する機会損失評価法を提案する。

2. 検索ログに対する概念グラフ

概念グラフでは，ある単語を含む文書集合がその単語の意味を規定するという考えに基づき，その文書集合の中での単語の上位下位関係を文書頻度を用いて定式化する。D を文書集合，w を単語とする。w が出現する D 中の文書数を $df(w, D)$ で表す。次に，単語 u, v の両方が表れる D 中の文書数を $df(u*v, D)$ で表す。単語 u と v が次の条件を満たすとき，u は v の上位であると定義する（$\alpha = 0.5$）。

$$\frac{df(u*v, D)}{df(v, D)} > \alpha, \quad df(u, D) > df(v, D)$$

つまり，単語 u と v について，v が現れる文書の過半数の文書で u が現れているときに，u は v より概念として上位にあるとする。

上で定義した単語の上位下位関係を，単語を点として上位下位関係を有向枝とする有向グラフで表すことにする。しかし，例えば，ある単語 u が単語の v の上位となっているとき，v の上位の単語がすべて u にとっても上位の単語となっているとする。このような単語が多い場合，生成された概念グラフは人間にとって分かりやすい表示になるとは限らない。このような場合は，u から v にのみ線を引く。つまり，各点に対しすぐ上の点を求め，「直上」の点との間だけに線を引く。これにより，各単語について局所的な上位下位関係にある単語を求めることができる。

検索連動広告では，広告出稿者が自社の商品やサービスに関連する適切なキーワードを選択することにより，検索エンジン利用者を効率良く自サイトに誘導することができる。この時，広告出稿者はどのようなキーワードに対して広告を出稿するかが問題となる。検索連動広告における「キーワードの発見」は「キーワード自体を発見すること」と「既にあるキーワードのリストから値段等を考慮して利益が最大になるようなキーワードを発見すること」の二段階に分けられる。「頭に浮かんでいないキーワードの発見」とは広告を出稿する際に候補として挙げられていないものを見つけることであり，「既にあるキーワードのリストから値段を見ての入札すべきキーワード発見」とは広告を出稿するキーワードの候補の中から，入札価格，クリック数等を計算し，目的とする利益やコストになるようにキーワードを選択することである。つまり，「頭に浮かんでいないキーワードを発見」することを繰り返しキーワード候補を作り，それを元に「既にあるキーワードのリストから値段を見て入札すべきキーワードを発見」し，広告の出稿が行われる。本稿で扱う「キーワードの発見」とは前者の「頭に浮かんでいないキーワードの発見」のことである。類義語や同義語も検索語として使われる。そのキーワードの代わりになるようなキーワードの発見を目的とする。また，検索者は複数の単語を組み合わせたクエリを用いて検索を行うため，出稿すべき単語の組み合わせも重要になってくる。本稿で提案するシステムでは，あるキーワードがあった際に，そのキーワードと共起すべきキーワードやそのキーワードの代わりになるようなキーワードの発見を目的とする。

例えば，美容用品を売っているサイトの検索連動広告への出稿候補となるキーワードが「美容」のみであったとする。このときに本稿で提案するキーワード発見支援システムが，「美容　クリーム」をカバーすることで「美容」に出稿することをある程度カバー出来ることを発見する。また同時に「美容」の代わりに「美肌」というキーワードが使えることを発見する。このように「美容　クリーム」や「美肌」などの今まで広告出稿の候補として挙げられていなかったものを発見することが本稿で提案するシステムの目的である。

「頭に浮かんでいないキーワードを発見」することは人間のひらめきに頼っている。そのようなキーワードを発見する際には，あるサイトに到着した際の検索クエリのような，キーワードが列挙されたものの中から選び出す手法が一般的である。また人のひらめきにより選び出されるため，キーワードを選ぶ際の

基準も曖昧である。

検索者の傾向をつかみ, かつ周辺語を見つけることが重要である。本稿で提案するシステムでは, 概念グラフを使って, 各キーワードに上下関係を定義し可視化することで, ただキーワードが列挙されている状態よりも, キーワード間の関連性を感覚的に掴みやすくすることが出来る。

本節ではある求人サイトに到着したときの検索クエリ (以下求人ログ) を対象とし概念グラフを生成した。この実験データは, 2006 年 4 月と 5 月に収集したものであり, 9,743 個の単語で構成される 15,874 個のクエリがあり, トータルでの訪問件数は 44,765 件である。表 1 は検索クエリで使われた出現頻度上位の単語を示している。

表 1　順位上位 10 個の検索クエリとキーワード

検索回数	検索クエリ	頻度	キーワード
2978	求人	14172	求人
168	医療事務　求人　正社員	3188	募集
153	サイト名	2794	アルバイト
125	トリマー　求人	1393	正社員
92	求人情報	1377	大阪
92	茨城　求人	1248	求人情報
79	求人　沖縄	1073	福岡
84	ミステリーショッパー　募集	1023	転職
80	駐車監視員　募集	808	東京
79	沖縄 求人	652	沖縄

図 2 は「求人」という単語が使われているクエリの集合から生成された概念グラフである。グラフでは, 出現頻度が 50 以上の単語を対象としている。図 2 から, 求人ログ中には「求人 地名」もしくは「求人 職種」という組み合わせが多く出現していることが分かる。このことから, 広告出稿者は求人というキーワードと共に, 職種と地名についても同時に入札すべきであるということが分かる。このように, 概念グラフを用いることにより, 出稿すべきキーワードの傾向が分かった。

図 3 は「募集」という単語が使われているクエリの集合から生成された概念グラフであり, 出現頻度が 20 以上の単語を対象としている。図 3 から, 「募集　駐車監視員」, 「募集　ミステリーショッパー (覆面調査員)」というクエリが多く出現していることが分かる。求人ログが取られたのは 2006 年 4 月, 5 月時点で, 駐車監視が民営化される直前であった。そのため「募集　駐車監視員」が注目されていたことが分かる。また駐車監視員だけでなくミステリーショッパー (覆面調査員) も, 同じく共起されている運転手や溶接工といった一般的な職種より興味を引いていることが分かる。つまりキーワード支援ツールにより「募集　"風変わりな職種"」という組み合わせが良いキーワードであることが想像できる。

対象サイトのクエリログの分析だけでは, 例えどんなに関連しているキーワードであっても, ログに含まれていないキーワードを発見することは出来ない。そこでキーワードアドバイスツールからキーワードを取得し, そのデータを元に概念グラフを生成しキーワードの発見を行うことにした。キーワードア

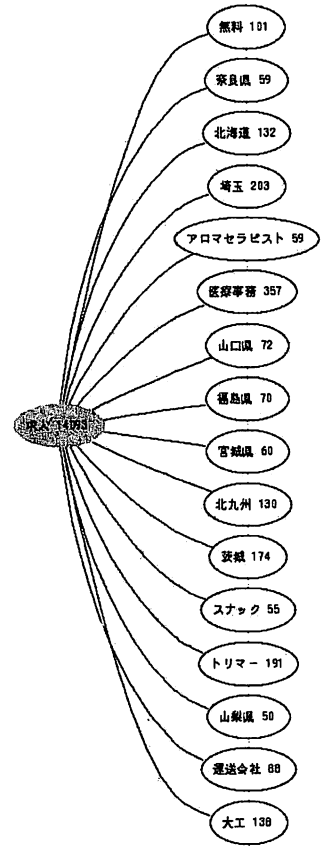

図 2　「求人」での概念グラフ

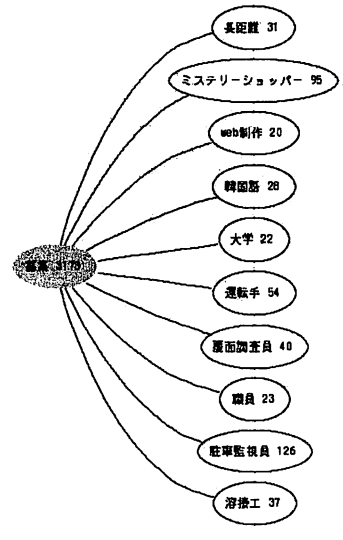

図 3　「募集」での概念グラフ

ドバイスツールは, 入力したキーワードを含む検索フレーズを, 検索された回数と共に列挙するものである。キーワードアドバイスツールにて列挙される検索フレーズは, 実際に Yahoo の検

索で入力されたクエリである。求人サイトのログ中に出現する各キーワードをそれぞれキーワードアドバイスツールへの入力キーワードとし，ログ中のキーワード上位 1000 個について世の中で投げられている関連キーワードを収集し，それらのキーワード群を元に概念グラフを生成する。キーワードアドバイスツールから得られたグラフと，サイトログから生成されたグラフを見比べることにより，ログには無いが世の中で投げられているキーワードの発見を行うことが出来る。

図 4 はキーワードアドバイスツールを使って「求人」という単語が使われているクエリの集合から生成された概念グラフである。色の付いているノードが，ログから作成される概念グラフでは上位下位関係を持たないが，キーワードアドバイスツールから作られる概念グラフでは上位下位関係を持つノードである。値段の付いているノードがログでの出現頻度上位 1000 個のキーワードである。グラフでは出現頻度が 200 以上の単語を対象としている。上位 1000 個に入っているキーワードは正社員だけで，他は取りこぼしていることが分かる。特にハローワークというキーワードは検索数が多いためこのキーワードに広告を出稿することが考えられる。

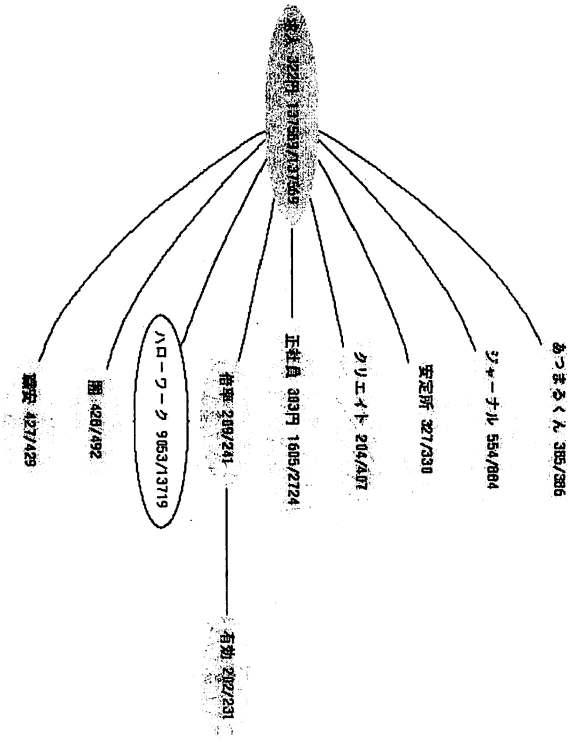

図 4 「求人」での概念グラフ

これまでの例についてはほとんどのグラフが深さ 1 となってしまい，深く関連を掘り下げることが出来なかった。概念グラフの定義における α を 0.5 よりも下げることでより多くの関連語が発見できることが確認できている。

例えば，「求人」という単語についても，共起率を $\alpha = 0.2$ に下げ，出現頻度を 1110 以上の単語を対象とすると深さが 4 となり，ログ中の上位 1000 個のキーワードも多く含まれるようになり，「情報　求人　アルバイト　短期」「情報　求人　技師　臨床」など今まででは分からなかったキーワードが発見できる。

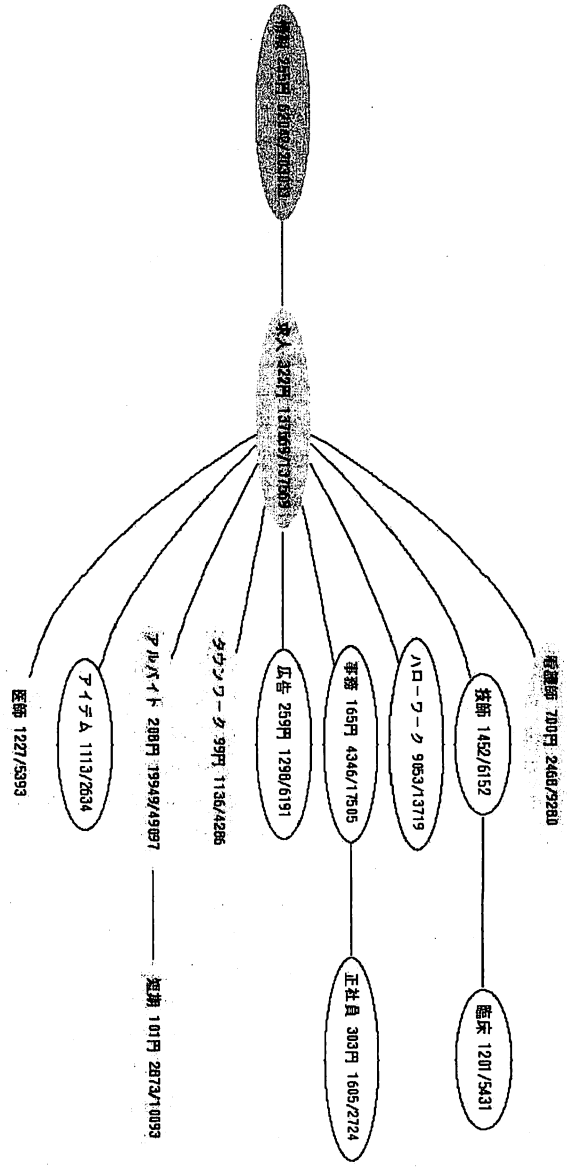

図 5 「求人」共起率 0.2 での概念グラフ

3. 大域共起率と局所共起率に基づく機会損失評価

これまでの手法では定量的なアドバイスが難しく，定性的なアドバイスしか出来なかった。共起率を利用することでフレーズの定量的なアドバイスを可能にする。フレーズとはキーワー

ドの組み合わせにより作られる検索クエリのこととする。キーワードアドバイスツールでの共起率とログでの共起率を見比べることでフレーズのアドバイスを行う。これによりどのフレーズが潜在顧客を逃しているかを定量的に判断することが出来る。

D をサイトログの文書集合, D' をアドバイスツールによる文書集合とすると, フレーズ「$u\ v$」について, サイト訪問者に対する共起率, ならびに検索エンジン利用者の大域的検索行動における共起率はそれぞれ, $LC(u,v) = df(u*v, D)/df(u, D)$ ならびに $GC(u,v) = df(u*v, D')/df(u, D')$。となる。顧客候補が全てサイトに訪れる場合に, 期待される訪問数が最大となるので, 機会損失の最大値は $df(u*v, D') - df(u*v, D)$ で求めることができる。また, ログの母集団の大きさを変えずサイト訪問者についての共起率を大域的共起率まで大きくできたときが, 機会損失が最小となり, その値は $df(u, D) * (GC(u,v) - LC(u,v))$ で求めることができる。

共起率がログ中でもアドバイスツール中でも出ているものが 5056 件あり, その内, ログ中の共起率が大きいものが 3734 件あった。これは 74%のフレーズについてログ中の共起率の方が大きいといえる。

共起率の差の大きい順に並べたものが表 2 である。一般的に一つの単語として使われているがキーワードアドバイスツールの形態素解析に合わせた結果生じているものが多い (ケアマネージャー, 行政独立, 専門員, 介護員, など)。「払い　週」や「製作　ホームページ」など求人に関係あるフレーズも上位に入ってきており, 定量的なアドバイスを可能にしている。

表 2　共起率の差が大きい順

フレーズ	サイト共起率	大域共起率	共起率の差
ケア　マネージャー	0.049	0.956	0.907
行政　独立	0.220	0.875	0.655
払い　週	0.085	0.658	0.573
員　専門	0.002	0.512	0.514
員　介護	0.023	0.029	0.510
製作　ホームページ	0.037	0.544	0.507
...	...	...	...

機会損失の最小値が大きい順に並べたものが表 3 である。「求人　情報」や「求人　アルバイト」, 「アルバイト　情報」など求人サイトに直接関係あるフレーズを定量的にアドバイス出来ている。

表 3　機会損失の最小値が大きい順

フレーズ ($u\ v$)	u の回数	共起率の差	機会損失の最小値
求人　情報	21491	0.361	7758.25
求人　アルバイト	21491	0.137	2944.27
県　市	5534	0.355	1964.57
アルバイト　情報	4250	0.342	1453.50
大阪　府	1920	0.347	666.24
神奈川　市	1288	0.497	640.17
...	...	...	...

4. まとめと今後の課題

検索連動広告は, 検索エンジンの検索結果に自サイトへのリンクを有料表示するサービスであり, 自社の商品やサービスに関連する適切なキーワードを選択することにより, 検索エンジン利用者を効率よく誘導することができる。検索連動広告の利用者の多くは, キーワードアドバイスツールから得られる情報をきっかけとしたひらめきによる試行錯誤でキーワードを決めなければならない。本稿では, 概念グラフの理論を検索ログに適用することにより, 検索連動広告における適切なキーワードを発見する方式を提案した。実データによる定性的な評価を行った。また, サイトログだけでなく, 検索エンジンにおけるクエリログを利用することにより機会損失の数量的評価が可能であることを示した。今回の報告は, 短期間のサイトログとアドバイスツールを使って得られる小規模なデータに基づく評価であるが, 有効性は検証できたと考えられる。より一般的な有効性評価のためには, サイトログも大域ログも, 長期的で大規模なデータについての分析と実証実験が必要である。

文　献

[1] Steven M. Beitzel, Eric C. Jensen, David D. Lewis, Abdur Chowdhury, Ophir Frieder, Automatic classification of Web queries using very large unlabeled query logs, ACM Transactions on Information Systems (TOIS), Volume 25 , Issue 2 (April 2007)

[2] A. Broder, A taxonomy of Web search. SIGIR Forum Vol.36, No.2 , pp.3–10, 2002.

[3] Bernard J. Jansen, Amanda Spink: How are we searching the World Wide Web? A comparison of nine search engine transaction logs. Information Processing and Management, 42(1), 248-263, 2006.

[4] E. Jensen, S. Beitzel, A. Chowdhury, O. Frideer, Query Phrase Suggestion from Topically Tagged Session Logs, In Proceedings of the Seventh International Conference on Flexible Query Answering Systems (FQAS 2006), Milan, Italy, June 2006.

[5] Ying Li, Zijian Zheng, Honghua Dai, KDD CUP-2005 report: facing a great challenge, ACM SIGKDD Explorations Newsletter, 7(2), pp.91 – 99, 2005.

[6] Viet Bang Nguyen and Min-Yen Kan, Functional Faceted Web Query Analysis, in: Query Log Analysis Social and Technological Challenges, A workshop at the 16th International World Wide Web Conference May 8, 2007 - Banff, Alberta, Canada

[7] M. Sanderson, S. Dumais, Examining repetition in user search behavior, Proc. ECIR 2007, pp. 597–604, 2007.

[8] 下司義寛, 和多大樹, 廣川佐千男, 英和辞典からの知識抽出, 第 68 回情報処理学会全国大会講演論文集 3, pp. 19–20, 2006.

[9] 滝井秀典, 1 億稼ぐ「検索キーワード」の見つけ方 儲けのネタが今すぐ見つかるネットマーケティング手法, PHP 研究所, 2006.

[10] Jaime Teevan, Eytan Adar, Rosie Jones and Michael Potts. History Repeats Itself: Repeat Queries in Yahoo's Logs. In Proceedings of the 29th Annual ACM Conference on Research and Development in Information Retrieval (SIGIR '06), Seattle, WA, August 2006.