検索結果中のアクセス集中サイトを利用したクエリ拡張法の提案

A Query Expansion Method Using Access Concentration Sites in Search Result

村田 眞哉 ♥ 戸田 浩之 ♦
松浦 由美子 ♠ 片岡 良治 ♣

Masaya MURATA Hiroyuki TODA
Yumiko MATSUURA Ryoji KATAOKA

近年、モバイルコンテンツ検索に対する有効なランキング手法の研究が盛んに行われている。有望な手法の一つにクエリ拡張が挙げられるが、実際のサーチエンジンでの利用を考えた場合には、計算コストの面から少ない拡張語数で精度を上げる事が必要とされている。そこで我々はクリックログに着目し、クリックが集中している検索結果のタイトルやスニペット (アクセス集中サイト) にはユーザが有用だと判断した表現があり、そこから取得されるキーワードがそのような拡張語の候補になるのではないかと考えた。そこでアクセス集中サイトをクリックログから的確に判別する方法を考案し、そこから取得される拡張語によるクエリ拡張法と、拡張語と初期のクエリに対する近傍性に基づく re-ranking を結び付けた新手法を提案し、実験により検証した。

Many studies are underway on ranking algorithms for mobile contents search. Query expansion is a quite promising method, but we need to achieve high search effectiveness with few expansion terms to reduce the calculation cost. Our solution is to focus on click logs, considering that there are expressions that users have found useful in the title and snippet of search results clicked intensively (access concentration sites) and these expressions become expansion term candidates. We introduce a technique that can precisely identify the access concentration sites using click logs and present a new method in which we combine a query expansion method using these terms with re-ranking based on the proximity of those terms with the original query. We verify our method in an experiment.

♥ 日本電信電話株式会社 NTT サイバーソリューション研究所 murata.masaya@lab.ntt.co.jp
♦ 正会員 日本電信電話株式会社 NTT サイバーソリューション研究所 toda.hiroyuki@lab.ntt.co.jp
♠ 日本電信電話株式会社 NTT サイバーソリューション研究所 matsuura.yumiko@lab.ntt.co.jp
♣ 正会員 日本電信電話株式会社 NTT サイバーソリューション研究所 kataoka.ryoji@lab.ntt.co.jp

1. はじめに

近年モバイル環境からの Web アクセスが増え、アクセス可能なモバイルサイトが爆発的に増加している。これに伴いモバイルサイトに対する検索サービスの需要が高まり、モバイル検索のランキング精度向上が重要な研究課題となっている。

しかしウェブサイトに対する従来の有効なランキング手法をそのままモバイルサイトに応用しても、例えば以下に挙げるモバイルサイト特有の性質により必ずしもよいランキング結果を得ることができない。

1. モバイルサイトは画面が狭いモバイル端末向けに作られているため、テキスト量がウェブサイトと比べ圧倒的に少ない
2. 有料の公式サイトや企業が運営しているサイトが多いことから、一般のウェブサイト群で見られるような有益な情報に対してリンクが張られているというよりは、サイト内をナビゲートする為のリンクが多い

1 は Okapi BM25[1] のような bag-of-words ranking function の困難を、2 はウェブサイト集合に重要度の順序構造を導入した PageRank[2] のような link-based ranking の困難を意味している。さらに追い撃ちをかけるのが、クエリのキーワード数の少なさである [3]。携帯端末の予測変換機能により、各キーワードのスペルはウェブサイト検索に比べて正確にはなるが、テンキーの打ちにくさによりキーワード数は少なくなる [4]。

そこで我々はユーザが検索結果を見て、実際にクリックするまでに至る行動を考えた。そして、ユーザは検索結果のタイトルやスニペットを見て、有用だと判断したらクリックするのではないかと仮定した。以降、本論文ではこのクリックが集中している検索結果のタイトルとスニペットのことをアクセス集中サイトと呼ぶ。しかしながらアクセス集中サイトを特定し、このサイトに高スコアを与えるだけの手法は正しくない。なぜならタイトルやスニペットが良いからといっても、そのサイト自体が必ずしも良いとは言えないからである。そこで我々はクエリ拡張 (query expansion) を利用することを考えた。

我々の仮定によると、クリックが集中している検索結果のタイトルやスニペットにはユーザが有用だと判断したキーワードがある。従ってこのキーワードを含むサイトがクエリに対する正解サイトである可能性が高いということになり、このキーワードでクエリ拡張を行うことで検索結果の精度向上が期待できる。このアイデアを実現するためには以下の課題を解決しなければならない。

- アクセス集中サイトの的確な特定
- 計算コストの面からできるだけ少ない拡張語 (1〜5 個) で大きな精度向上を得られること

これら課題に対して我々はクリックログを解析することでアクセス集中サイトを的確に特定し、そのタイトルとスニペットから拡張語を取得する方法を考案した。そしてこの拡張語によるクエリ拡張と、拡張語とクエリに対する近傍性に基づく検索結果の

re-ranking を結び付けた新手法を提案し、計算機実験により検証した。この結果を報告する。

以下、2 章で関連研究、3 章で本論文に用いたベースライン、4 章で我々の提案手法について説明し、5 章でその実験結果を検証する。そして 6 章で本論文をまとめる。

2. 関連研究

クリックログを利用する一般的なクエリ拡張は、まず入力されたクエリとクリックログを照合し clicked sites を特定することから始まる。そしてこれを基に拡張語を取得する。Cui らはクリックログから clicked sites を特定し、その中の各ワードに対してクエリとの共起確率を計算し、この値の高い拡張語を用いてクエリ拡張を行っている [5]。しかしながら彼らの結果によると、拡張語 30〜50 個で最大精度を出しており、実用面から見ると計算コストがかかってしまう。

Shen らはクエリの変更過程 (re-formulation) にも注目している [6]。ユーザはクエリを入力し検索結果を眺め、もしこれが希望通りでなかったらクエリをより良いものに変更すると考えられる。このクエリの re-formulation はクリックログから取得可能で、彼らはこの過去一連の re-formulation を効果的に取り扱うことで現在のクエリを補完し、検索精度の向上を実現している。

Zhuang ら、Parikh らもクエリの re-formulation を利用する同様な手法を考案しているが、Shen らと異なるのはクエリの re-formulation をクエリ拡張ではなく、検索結果に対する re-ranking に用いている点である [7] [8]。

3. ベースライン

1. 我々が採用したベースライン 1 は文章長で正規化した $tf \cdot idf$ 法である。クエリ q が形態素 t_i, $(i = 1, 2, 3, \cdots, n)$ から成るとし、ドキュメントを d_j と表す。このドキュメント d_j のスコア値は

$$score(d_j, q) = \sum_{i=1} \frac{log(1 + tf(t_i, d_j)) \times idf(t_i)}{log\ length_j} \quad (1)$$

で与えられる。ここで $length_j$ はドキュメント d_j の文章長を表す。そしてこのスコア値の高いサイトから順に並べ、これを検索結果とする。

2. ベースライン 2 は疑似フィードバック (pseudo-feedback) によるクエリ拡張である。拡張語の取得先としてクエリに対する検索結果上位 10 件を使用する。検索結果のタイトルとスニペットを形態素 t_i, $(i = 1, 2, 3, \cdots, n)$ に分解し、その $log(1 + tf(t_i, d_j)) \cdot idf(t_i)$ が高いものから順に拡張語として採用し、クエリ拡張を実行する。

4. 提案手法

提案手法はクリックログを用いたクエリ拡張がベースであるが、従来との違いは以下にある。

- アクセス集中サイトを用いたクエリ拡張
- 拡張語とクエリとの近傍性に基づく検索結果の re-ranking

4.1 アクセス集中サイトを用いたクエリ拡張

我々は数多くのユーザが有用だと判断し、クリックが集中しているサイト (アクセス集中サイト) から拡張語を的確に取得したい。そこでクリックログを解析し、検索結果のランク vs クリック回数の座標上でグラフを描き、一般的な傾向である指数関数的減少曲線と比べて傾きが大きく変化しているサイトに注目する。これは通常とは異なるサイトがそこに存在することを意味しており、その一つ後ランクのサイトのクリック回数が相対的に大きく減少しているサイトがアクセス集中サイトであると仮定した。

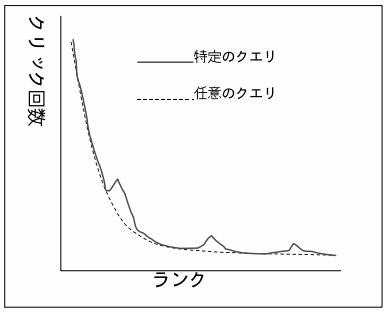

図 1 ランク vs クリック回数の曲線例。上に凸の山ができているランクのサイトにアクセスが鋭く集中していると考える。ゆえにその一つ後ろのサイトは通常と比べて傾きが大きく減少する

Fig. 1 A curve example of click number versus results rank. We consider sites that demonstrate strongly-peaked curves to be intensively accessed. The inclination to the next rank site considerably decreases compared with a normal curve.

なぜならユーザに有用だと判断されたサイトは数多くクリックされ、この座標上に上に凸の山ができると考えられるからである。

図 1 を例にとり、この考えを詳しく説明する。横軸が検索結果のランク、縦軸がクリック回数である。クリックログに対して全クエリを用いて解析し、描いた曲線を任意クエリ曲線、クエリを一つ固定して描かれた曲線を特定クエリ曲線と呼ぶ。もしこの図のように、任意クエリ曲線に対して特定クエリ曲線が、あるランク箇所に上に凸の山を描いているならば、このランクのサイトにアクセスが鋭く集中していると考える。これは極端な例ではあるが、本提案手法はこのアクセス集中度合の微妙な変化を、特定クエリ曲線のランク r とランク $r+1$ における傾きで捉える。

クエリに対するあるサイトのランク $r_{q,j}$ とその一つ後のランク $r_{q,j}+1$ のサイトに注目し、それぞれのクリック回数を $cc_1(q, r_{q,j})$、$cc_1(q, r_{q,j}+1)$ とする。また全クエリに対する検索結果のランク r_j と r_j+1 のクリック回数を $Tcc_2(r_j)$、$Tcc_2(r_j+1)$ とすると、

$$\begin{aligned} Inc(q, d_j) = &\arctan\left((cc_1(q, r_{q,j}+1) - cc_1(q, r_{q,j}))/1\right) \\ &+ \arctan\left((Tcc_2(r_j+1) - Tcc_2(r_j))/1\right) \end{aligned}$$

の負値が大きいサイトがアクセス集中サイトで、この絶対値がアクセス集中度合であると言える。なぜならアクセス集中サイトの次のランクのサイトは、相対的にクリック回数が大きく減少し、集中度合の高いサイトほどこの減少量が大きいと考えられるからである。$arctan(x)$ は傾きを角度に変換する関数であり、これを用いる理由を以下に述べる。一般的に上位ランクのサイトほど数多くクリックされる傾向があるので、上位ランクの傾きは下位ランクと比べて簡単に大きくなり、これでは正確にアクセス集中サイトを抽出することができない。そこで約 $x = 2$ 以降で急激に値の増加が減少する $arctan(x)$ の性質を利用し、このバイアスを軽減した。結果、値域は $-1.5 \simeq Inc(q, d_j) \simeq 1.5$ になる。第 2 項目は下位ランクのクリック偶発性に対処する項である。そしてこの $Inc(q, d_j)$ が -2.0 より小さいランクのサイトをアクセス集中サイトとみなし、そこから拡張語群を取得した。

拡張語の順序付けは、$Inc(q, d_j)$ の絶対値に拡張語の IDF を掛けた

$$w(t_i) = \sum_{d_j} idf(t_i) \times log(1 + |Inc(q, d_j)|) \qquad (2)$$

の値で行い、この上位 N 個でクエリ拡張を実行し、式 (1) に基づき検索を行う。

4.2 拡張語のクエリに対する近傍性に基づく re-ranking

4.1 節の方法で得られる拡張語はドキュメント中のクエリに対し、近い距離 (近傍) にあることでさらなる効果を生むと期待できる。なぜならボディに対する拡張語はアクセス集中サイトのスニペットから取得しており、我々の検索システムのスニペットはクエリを中心とする決まった形態素数分を表示するからである。ゆえに本手法を我々のような検索システムに適用すると、拡張語を取得する段階において、自動で拡張語にクエリに対する近接性を含ませていることになるからである。これを踏まえ、4.1 節のクエリ拡張で得た検索結果上位 100 件をさらに re-ranking することを考える[1]。

クエリを中心とする形態素数 25 個分の窓 L を検索結果の各ドキュメントボディから抜き出し、この中に拡張語 t_i が含まれていればそれに対応する値 (式 2) を加算していく。窓 L に $n(t_i)$ 個の拡張語 t_i が含まれる場合の re-ranking ドキュメントスコア値 $r\_score(d_j, q)$ は、最初の検索結果のドキュメントスコア値を $score(d_j, q)$ とすると、

$$r\_score(d_j, q) = score(d_j, q) + \sum_{t_i \in L_j} n(t_i) \times w(t_i) \qquad (3)$$

で与えられる。このスコアの大きいサイトから順に並べ、これを最終検索結果とする。この re-ranking までの一連の処理が我々の提案手法である。

5. 実験結果

実験結果を図 2 に示す。

[1] 近接検索でもよい

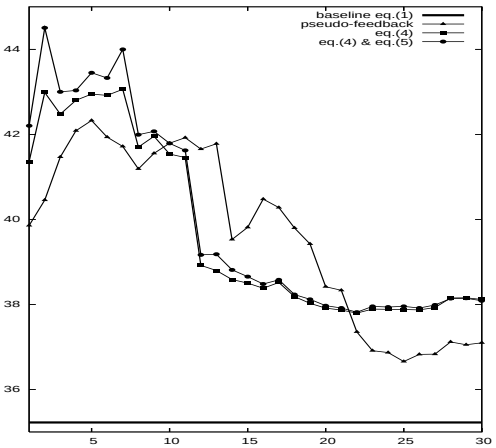

図 2 縦軸は MAP、横軸は拡張語の数。直線は 3 節のベースライン 1 で、▲がベースライン 2 の疑似フィードバックによるクエリ拡張、■がアクセス集中サイトを用いたクエリ拡張、●がさらに■を拡張語のクエリに対する近傍性に基づき re-ranking した実験結果を表す

Fig. 2 Vertical axis is MAP and horizontal axis is the number of expansion terms. Straight line plots our baseline 1 at Section 3, ▲ line plots the query expansion with pseudo-feedback(baseline 2), ■ line plots the query expansion with access concentration sites, ● plots the re-ranking of ■ based on the proximity of expansion terms with the original query.

結果はベースラインの MAP が 35.2% であるのに対して、予備実験の結果最も精度の高かった疑似フィードバックによるクエリ拡張は拡張語数 1〜5 個で平均 MAP41.23% である。式 (2) に基づくアクセス集中サイトを用いたクエリ拡張では拡張語数 1〜5 個で平均 MAP42.51%、さらに式 (3) によりこの検索結果を re-ranking する我々の提案手法では拡張語数 1〜5 個で平均 MAP が 43.23% になっている。また最大 MAP は拡張語数 2 個で 44.5% に達している。これはベースラインの精度を大幅に改善しており、本提案手法の有効性を実証している。この結果はアクセス集中サイトから得られる拡張語がクエリに対する highly relevant terms であることを意味しており、クエリに対する近傍性と結びつけることによって、少ない拡張語数 (1〜5 個) でのクエリ拡張の精度をさらに向上させる効果があることを示している。精度が向上したクエリ群に明確な共通性は無いが、固有名詞 (人名、店名、商品名) などが比較的多く含まれていた。表 1、2 に有効であったクエリの一部例を、title、snippet それぞれから取得された拡張語と共に記す。

6. まとめ

以上、本論文ではアクセス集中サイトを多くの検索結果の中からクリックログを解析することにより判別し、そこから取得される拡張語を用いてクエリ拡張を行い、拡張語のクエリに対する近傍性に基づき検索結果を re-ranking する新手法を提案し、モバ

表 1 「あかひげ」に対して抽出された拡張語の例
Table 1 An example of expansion terms for “akahige”

	傾き正規化	疑似フィードバック
1 位	シソ酢 (from title)、 薬局 (from snippet)	akahige(from title)、 全国温泉 (from snippet)
2 位	薬局、NEW	あか、あか
3 位	あか、新橋	薬局、泊まれる

表 2 「トイザらス」に対して抽出された拡張語の例
Table 2 An example of expansion terms for “toizarasu”

	傾き正規化	疑似フィードバック
1 位	日本トイザらス、 日本トイザらス	BLYTHEROOM、 PR
2 位	Mobile、遊ベル	トイザらス DC、マイトカイザー
3 位	Bee、PR	Plaza、ネタなべ

イルコンテンツに対して実験を行った。この結果により、本論文の目的であった、少ない拡張語数 (1〜5 個) でベースラインを大きく上回る精度を得ることを達成し、本提案手法の有効性を実証した。

我々の提案手法は、数多くのユーザが有用だと判断したアクセス集中サイトから拡張語を的確に取得し、これに基づきクエリ拡張を行うことが重要であるという考えの下に成り立っている。そこで一つのクエリを固定してクリックログを解析し、検索結果ランク vs クリック回数の座標上でこの曲線を描き、傾きに注目した。数多くのユーザが有用だと判断し、クリックが集中するサイトのランクと次のランクの間のクリック回数には大きな差ができるはずである。なぜならこのようなサイトは座標上に上に凸の山を描くと考えられるからである。そしてこの山の傾きの減少度合が大きくなればなる程、つまりこの山の底辺に対する傾きの角度が大きくなればなる程、このランクのサイトが数多くのユーザを引き付け、有用だと判断された証拠になると考えたのである。

そしてもう一つ重要なことは、拡張語のクエリに対する近傍性である。今回利用した検索システムのスニペットはクエリを中心とする決まった形態素数分を表示する。ゆえに我々の手法をこのような検索システムに適用し取得した拡張語は、クエリの近傍にあればよりいっそうの効果があると期待できる。この考えに基づきクエリ拡張で得た検索結果を、拡張語のクエリに対する近傍性に基づき re-ranking した。

今後は本手法をさらに改善し、正例による正のフィードバックのみならず、負例による負のフィードバックも検討したい。そしてモバイルならではの問題と特徴を見極め、より多種多様なテストコレクションで実験を行う。

[文献]

[1] Stephen E. Robertson, Steve Walker, Susan Jones, Micheline Hancock-Beaulieu, and Mike Gatford, “Okapi at TREC-3”. Proceedings of the Third Text REtrieval Conference(TREC 1994)

[2] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, “The PageRank Citation Ranking: Bringing Order to the Web”. 1998

[3] Masaya Murata, Hiroyuki Toda, Yumiko Matsuura and Ryoji Kataoka, “A Query Expansion Method Using Access Concentration Sites in Search Result”. Proceedings of the DataBase and Web symposium (DBWeb 2007)

[4] 佐野正弘, ” 大人が知らない携帯サイトの世界〜PC とは全く違うもう 1 つのネット文化〜” マイコミ新書, 2007

[5] Hang Cui, Ji-Rong Wen, Jian-Yun Nie and Wei-Ying Ma, “Probabilistic Query Expansion Using Quer Logs”. Proceedings of the 11th international conference on World Wide Web 2002, 325-332

[6] Xuehua Shen, Bin Tan and ChengXiang Zhai, “Context-Sensitive Information Retrieval Using Implicit Feedback”. Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval 2005, 43-50

[7] Ziming Zhuang and Silviu Cucerzan : “Re-Ranking Search Results UsingQuery Logs”. Proceedings of the 15th ACM international conference

[8] Jignashu Parikh and Shyam Kapur, “Unity: Relevance Feedback using User Query Logs”. Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval 2006, 689-690

村田 眞哉 Masaya MURATA
NTT サイバーソリューション研究所 所属．2007 年早稲田大学大学院理工学研究科修了．同年, 日本電信電話 (株) 入社．情報検索の研究に従事．情報処理学会正会員．

戸田 浩之 Hiroyuki TODA
NTT サイバーソリューション研究所 所属．1999 年名古屋大学大学院工学研究科博士前期課程修了．2007 年筑波大学大学院システム情報工学研究科博士後期課程修了．1999 年日本電信電話 (株) 入社．以来、情報検索、情報抽出の研究に従事．博士 (工学)．ACM SIGIR, 情報処理学会, 電子情報通信学会会員．

松浦 由美子 Yumiko MATSUURA
NTT サイバーソリューション研究所 主任研究員．1993 年慶応義塾大学大学院理工学研究科修了．同年、日本電信電話 (株) 入社．マルチメディアの研究に従事．情報処理学会会員．

片岡 良治 Ryoji KATAOKA
NTT サイバーソリューション研究所 主幹研究員．1987 年千葉大学大学院電子工学専攻修士課程修了．同年、日本電信電話 (株) 入社．トランザクションの並行処理制御方式の研究、マルチメディア情報システムの研究、ポータルサービスシステムの研究開発に従事．情報処理学会会員の研究に従事．情報処理学会正会員．