

検索ログを用いた検索クエリの困難度判定によるクエリ推薦方式

A Method that recommend a query by difficulty judgment using retrieval data

宮原 伸二 数原 良彦 小長井 俊介 片岡 良治
Shinji Miyahara Yoshihiko Suhara Shunsuke Konagai Ryoji Kataoka

NTT サイバーソリューション研究所
NTT Cyber Solution Laboratories

Users can use a query, and came to be able to access various pages in the Internet in recent years. When the user retrieves the page, it exists a query that the retrieval is easy and difficult query. We propose the method that recommend a query by difficulty judgment using retrieval data.

1. はじめに

近年、検索ポータルサイトにおける検索ログやクリックログを利用し、ユーザの検索支援に役立てる取り組みが注目されている。特に、ユーザの検索活動を効率化するための支援として、ユーザの入力するクエリに対し、他のユーザが頻繁に利用する関連したクエリを推薦したり、絞り込みのクエリを推薦する試みが広く行われている。この推薦されるクエリによって、ユーザは欲しい Web ページに対する効果的なクエリを得ることができる。それにより、欲しい Web ページに対するコンテキストが少ないユーザでも、効率的に検索活動が行える。

一般に、検索ログを用いたクエリ推薦では、検索ログにおける URL のクリック数である CF(Click Frequency)やクエリの利用頻度である QF(Query Frequency)を用いて、クエリと URL 間の結びつきを確率的に求める研究が行われている[Deng 2009]。このクエリと URL 間の確率により、頻繁にクリックされる URL 集合にアクセスするための効果的なクエリを抽出できる。また、URL を介したクエリ同士における到着時間を用いたクエリ推薦に Hitting Time を利用した手法がある[Mei 2008]。この手法では、クエリ間の Hitting Time を再帰的に計算し、得られた Hitting Time の小さいクエリ同士に強い関連があると考えている。この手法では、入力したクエリで得られる URL 集合と関連が強い URL 集合に結びつくクエリが推薦される。

これら従来のクエリ推薦手法では、頻繁に利用されるクエリや入力したクエリに結びつきの強い関連したクエリが推薦されるが、推薦されたクエリによる検索結果の精度が考慮されていない。そのため、推薦されたクエリを用いて、ユーザの欲しい Web ページが得られないシーンが存在する問題がある。

そこで本研究では、ユーザの欲しいページを容易に見つけられるクエリと推薦するため、検索ログを用いてクエリの困難度を数値化し、その困難度に応じたクエリを推薦する方式を提案する。さらに、クエリ困難度と検索結果の人手評価による困難度を比較するとともに、従来手法と提案手法に対し評価実験での優位性を実証する。

2. 検索クエリの困難度

一般に、ユーザの入力するクエリには欲しいページが見つかりやすいクエリと、見つかりにくいクエリが存在する。これら見つかりやすいクエリと見つかりにくいクエリを図 1 を用いて説明する。

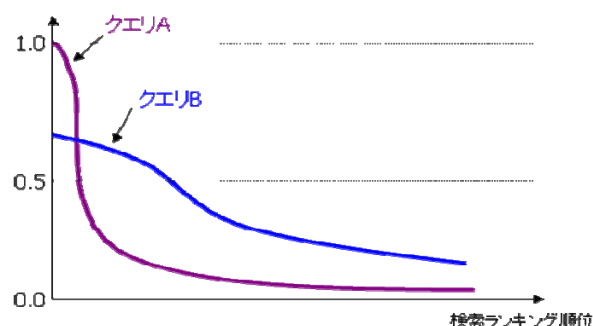


図 1: クリックされた URL と順位の関係

図 1 では、縦軸に検索でクリックされた URL の数を入力されたクエリの数で割った URL の数値を表している。横軸は、URL の検索ランキング順位を示している。クエリ A は、ほとんどのユーザが検索結果上位の URL をクリックし、他の URL をクリックせずに検索を終えている例である。この例において、ユーザは検索結果の上位に欲しいページが存在し、そのページで満足して検索を終えていると考えられる。次に、クエリ B は、検索結果上位の URL でもユーザにクリックされないケースが多く存在し、かつ、検索結果下位の URL もまんべんなくクリックされている。この例において、ユーザは検索結果上位に欲しいページが存在せず、検索結果下位からページを探すためにクリックしていると考えられる。そこで我々は、ユーザの欲しい Web ページが見つかるクエリの困難度を以下のように定義する。

$$\text{困難度} = \text{平均クリック順位} \times \sum \{-1.0 \times P(q) \times \log(P(q))\} \quad (1)$$

式(1)の第一要素は、ユーザがクリックした URL の検索ランキング順位の平均値が高いほど、検索結果の下位の URL をクリックしているため欲しいページが見つからないと考える。また、式(1)の第二要素はクリックによるエントロピーを表しており、エントロピーが大きいほど、ユーザがクリックする URL が特定の URL に集中していないため、ユーザは欲しい Web ページが見つからないと考える。

連絡先: 宮原 伸二, NTT サイバーソリューション研究所, 横須賀市光の丘 1-1, miyahara.shinji@lab.ntt.co.jp, Tel: 046-859-4923, Fax: 046-855-1730

3. 困難度を考慮したクエリ推薦

我々は、前述のクエリの困難度を基に、ユーザの入力したクエリに関連するクエリの中から、困難度の小さいクエリを推薦してユーザの検索活動を支援するクエリ推薦方式を提案する。提案手法では、クエリの困難度を算出し、クエリと URL から構成される二部グラフから関連するクエリと URL のサブグラフを構築して推薦するクエリを抽出する。

まず始めに、検索ログを基に検索ログ内に出現するクエリに対し、検索クエリの困難度を算出する。困難度の算出には、式(1)を用いて算出する。

次に、検索ログで出現するクエリとクリックされた URL に対する二部グラフを作成する。この二部グラフでは、入力されたクエリとクリックされた URL をエッジで連結したものである。この二部グラフを基に、あるクエリに着目し、着目したクエリを出発点として、深さ優先探索で予め設定した閾値の深さまで探索し、探索で得られ二部グラフに関連するクエリのサブグラフとする。

ユーザが入力したクエリに対し、入力したクエリを含むサブグラフを抽出し、そのサブグラフ内に存在するクエリの中から、困難度の最も小さいクエリを N 件抽出して、それらクエリを困難度の小さい順にソートしたものをユーザに推薦するクエリとする。

4. 評価実験

本研究では、提案手法の有効性を確認するため、検索ログと検索結果に対する人手評価データを用いて評価実験を行った。検索ログは商用の PC 検索における、入力されたクエリとクリックされた URL がついでとなっている約 1000 万行の検索ログを用いた。人手評価は、ログ内に出現する約 500 件のクエリに対し、クエリの検索上位 200 件の URL に人手評価値を付与した。ここで、検索結果の URL は 6 段階で評価し、評価値 6 は最もクエリに適合し、評価値 1 は最もクエリに適合しないものとしている。

評価実験では、クエリの困難度と検索結果に対する評価値の傾向分析と、従来手法と提案手法の比較を行った。

4.1 人手評価データを用いたクエリ困難度の傾向

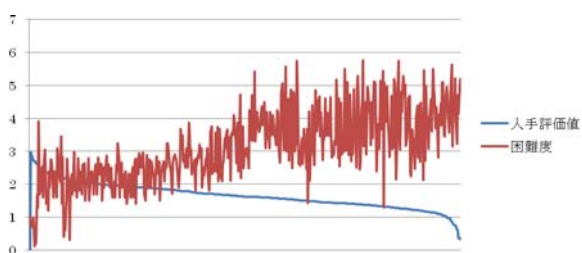


図 2: クエリ困難度と人手評価値の関係

我々は、提案するクエリ困難度と、人手によるクエリの検索結果の精度の関係を調査した。図 2 にクエリの困難度と、クエリに対する検索結果の評価値の平均のグラフを示す。グラフの縦軸はクエリの困難度と検索結果の人手評価による平均値を表している。また、横軸は検索結果の平均値でソートしたクエリを表している。

このグラフから、提案するクエリ困難度が大きいほど、検索結果の評価値が小さくなる傾向を示している。このことから、提案するクエリの困難度は、ユーザにとって欲しいページが見つかり難さを数値化できることに役立つと考えられる。

CF-IQF	Hitting Time	提案手法
NTT 東日本 フレッツ(2.6)	光テレビ(1.8)	フレッツ NTT(2.9)
光プロバイダー(1.4)	NTT 光フレッツ(2.3)	NTT 東日本 フレッツ(2.6)
ADSL(1.3)	フレッツ NTT(2.9)	NTT B フレッツ(2.9)
NTT 東日本 フレッツ(2.6)	B フレッツプラン(2.1)	フレッツ光(2.9)
NTT 光フレッツ(2.9)	フレッツ光(2.9)	光テレビ(1.8)

表 1: 各手法で推薦されるクエリ

4.2 提案手法と従来手法の比較

提案手法で推薦されるクエリと、従来手法で推薦されるクエリを比較した。従来手法では、2 つの手法を用いた。一つは、CF-IQF を用いて再帰的にクエリと URL 間のエッジの重みを計算してクエリ間の関連度を用いた。二つ目は Hitting Time を用いてクエリ間の関連度を算出した。表 1 に、クエリ「B フレッツ」に対して、各手法で推薦されるクエリの一部の例を示す。表 1 の括弧内は人手評価値の平均値である。

5. 考察

前述の 2 つの評価実験に関する考察を行う。クエリ困難度と検索結果の平均評価値から、クエリ困難度が検索結果の精度を表していることを確認した。このことから、クエリ推薦の際に、クエリ困難度の低い関連クエリを推薦することで、ユーザの検索活動が効率化できると考えられる。

次に、提案手法と従来手法による推薦クエリを比較し、提案手法が検索精度の良いクエリを推薦できることを確認した。従来手法は、クエリの利用回数やクリック回数によるクエリ、URL 間のエッジの重みを重視しているため、利用頻度の高いクエリが推薦される傾向があるが、提案手法は利用回数が少なくても精度の良い検索結果が得られるクエリを推薦できる。提案手法としてのデメリットとして、ユーザが入力したクエリと推薦するクエリの一般的な関連度が低い場合も存在するため、推薦されるクエリをユーザが理解しづらい可能性もある。

6. おわりに

本研究では、検索ログを用いたクエリ推薦において、推薦したクエリによる検索結果の精度に着目し、クエリに対する検索結果の困難度を算出する方式を提案した。また、このクエリの困難度を用いたクエリ推薦方式を提案した。評価実験では、提案するクエリの困難度と、人手による検索結果の評価値の関係を調査し、クエリの困難度が検索結果の精度を表す傾向となることを確認した。さらに、提案するクエリ推薦手法と従来手法を比較し、提案手法で推薦するクエリによって精度の良い検索結果が得られることを確認した。

今後は、クエリの困難度と検索ログに含まれるクエリとクリックされた URL の二部グラフを基に、グラフ全体の特徴を考慮したクエリ推薦手法のアルゴリズムに取り組み、評価実験による手法評価を実施する。

7. 参考文献

- [Mei 2008] Qiaozhu Mei, Dngyong Zhou, Kenneth Church: Query suggestion using hitting time, In Proceedings of CIKM '08, pages 469-478, ACM, 2008.
- [Deng 2009] Hongbo Deng, Irwin King, Michael R. Lyu, Entropy-biased Models for Query Representation on the Click Graph, In Proceedings of SIGIR '09, pages 339-346, ACM, 2009.