

検索システムユーザの分野別の知識推定

佐藤 大祐[†] 安田 宜仁[†] 望月 崇由[‡]

鈴木 智也[‡] 松浦 由美子[†] 片岡 良治[†]

[†] 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 〒239-0847 神奈川県横須賀市光の丘 1-1

[‡] NTT レゾナント株式会社 〒108-0023 東京都港区芝浦 3-4-1

E-mail: [†] sato.d@lab.ntt.co.jp

あらまし 検索システムを利用するユーザの、分野ごとの詳しさを推定する手法を提案する。提案法は、分野内におけるクエリの検索頻度に着目する。51名の被験者がスポーツ分野に関して検索した際のクエリログを用いて、知識量推定の評価実験を行った。実験の結果、提案手法による推定結果と、人手による推定結果との相関係数は0.67となり、手法の有効性を確認することができた。

キーワード 情報検索, クエリログ分析, パーソナライゼーション

Estimation of the User's Domain Knowledge in Information Retrieval System

Daisuke SATO[†] Norihito YASUDA[†] Takayoshi MOCHIZUKI[‡]

Tomoya SUZUKI[‡] Yumiko MATSUURA[†] Ryoji KATAOKA[†]

[†] NTT Cyber Solutions Laboratories, NTT Corporation 1-1 Hikarinooka, Yokosuka-shi, Kanagawa, 239-0847 Japan

[‡] NTT Resonant Inc. 3-4-1 Shibaura, Minato-ku, Tokyo, 108-0023 Japan

E-mail: [†] sato.d@lab.ntt.co.jp

Abstract We propose a method to estimate the domain knowledge of the user of search engines. Our method focuses on the frequency of each query in a specific domain. We conduct an evaluation experiment for knowledge estimation using query logs of 51 subjects who search in sports. Correlation coefficient between estimated results by the proposed method and human judgments was 0.67, which implies the effectiveness of our method.

Keyword Information Retrieval, Query log analysis, Personalization

1. はじめに

近年、検索システムを用いてインターネット上の情報を検索することは一般的になっており、多くの人々が検索システムを利用するようになっている。インターネット上で発信される情報は増加を続けており、膨大な量の情報の中から多様なユーザそれぞれに適合する情報を提示することが検索システムに求められている。

検索システムを用いた文書の検索結果には様々な内容の文書が含まれている。例えば同じ主題を扱った文書であっても、内容を理解するために事前知識を多く必要とする文書とそうでない文書が存在する場合がある。主題について多くの事前知識が必要とされる文書では、文書中に多くの専門用語が使用され、主題について事前知識をあまり必要としない文書では、使用

される専門用語について文書中で解説を行っている、または専門用語ではなくより平易な語を使って書かれていると考えられる。

ユーザが検索システムを用いて目的とする文書を得ることをサポートするために、専門的な知識を持っているユーザに対しては、主題について専門的に書かれている情報を、知識を持たないユーザに対しては、主題に対して平易な言葉で書かれている情報を提示することは効果的であると考えられる。

上記のような検索システムを実現するためには、ユーザの知識量を知り、情報を得るために必要となる事前知識を知り、ユーザの知識背景に見合った情報を提示する仕組みが必要であると考えられる。本稿ではこのうち、ユーザの分野に関する事前知識の量を推定する手法を提案する。手法により、このユーザは少なく

ともこの分野に関してはある程度事前知識を持っており、平易な言葉による解説は必要ないだろうということの提示を目的とする。

2. 関連研究

2.1. 文書の難易度

Nakatani ら^[1]により、クエリに対して専門知識を持たない一般的なユーザに対し、理解が容易である文書を検索結果の上位に提示する手法が提案されている。文書中の専門用語の出現密度と、文書の読みやすさの指標であるリーダビリティから、一般的なユーザにとっての理解の容易さを文書ごとに測定し、測定結果を理解容易度として文書に与えている。西原ら^[2]は、なじみのない言葉の意味を検索結果から理解するための支援システムを提案している。支援システムは文書に難易度を与え、さらに学習順序をユーザに提示するものである。これらの手法はともに文書の難易度を判定し、専門知識を持たない一般ユーザにとって理解しやすい文書を提示するものであり、ユーザの専門性の推定を行うという点で我々の提案手法と異なる。

2.2. ユーザプロファイリング技術

全てのユーザに対し同じ結果を提示するのではなく、ユーザそれぞれの背景を考慮に入れるアプローチとして、パーソナライゼーションがある。土方^[3]によると、個人に適した情報を選択し、個人に適した方法で表示するサービスをパーソナライゼーションといい、パーソナライゼーションを行うために必要となる、ユーザに関するプロフィール情報を抽出する技術をユーザプロファイリング技術という。ユーザプロファイリング技術として、ユーザの興味関心を抽出する手法等が提案されている^[5]。提案手法は、ユーザの分野ごとの知識というプロフィール情報を得ようとするものである。

3. 提案手法

ある分野においてその人が専門的であるということとは、その分野における事柄について多くのことを知っているということとする。また、分野に関する知識を蓄えていく過程では、分野内の事柄に関して、認知度の高い事柄からだんだんと認知度の低い事柄について知識を得ることから、ある分野内の事柄の中でも認知度の低い事柄について知っている人は、分野について多くのことを知っているだろうということが推測される。つまり、ある分野において認知度の低い事柄を知っているということから、分野において専門的であるということが推測される。

検索システムを利用して情報を得ようとする場合、

情報要求をクエリとしてシステムに投入する。情報要求を行っている事柄に関してユーザが知識を持っているということはいえないが、認知度の低い事柄について情報要求を行うことは、その分野に関してある程度の知識を有しているのではないかと考える。提案手法は、ある分野において、クエリとして入力される頻度の低い語句を使って検索するユーザは、その分野において高い専門性を持っているとの仮説に基づいている。例えば、「野球」という分野において、「ジャイアンツ」や「イチロー」といった、入力される頻度の高いクエリを入力するユーザよりも、「クイックモーション」や「ツーシーム」といった、入力頻度の低い語句をクエリとして選んでいるユーザのほうが、「野球」について多くの知識をもっているのではないかと推測される。

各分野においてのユーザの専門性を調べるため、クエリの専門性の指標として以下に定める IQF (Inverse Query Frequency) を用いる。以下に、分野 d におけるクエリ t の IQF を表す $iqf_d(t)$ を求める式を示す。

$$iqf_d(t) = \log \frac{N_d}{qf(t)} + 1$$

ここで N_d は、クエリログ中の対象分野 d に属する全クエリの数、 $qf(t)$ は、クエリ t を入力したユニークユーザ数である。クエリ t が分野 d に対して専門性の高い語句であるほど $iqf_d(t)$ が高くなる。

$iqf_d(t)$ からユーザの分野における知識を推定する。ユーザ u の、分野 d に属する語句全ての IQF の集合を $Q_{u,d}$ として、IQF に基づくユーザのある分野における知識量推定値 $C(Q_{u,d})$ を求める。

$$C(Q_{u,d}) = \frac{\sum_t iqf_d(t)}{N_{u,d}}$$

ここで $N_{u,d}$ は、ユーザ u が入力したクエリのうち、対象分野 d に属する語句がクエリとして入力された回数である。

4. 予備実験

提案手法を適用する分野として適切なものを調べるため、予備実験を行った。国内の商用検索システムの 2009 年 6 月 1 日～6 月 10 日までの約 2000 万クエリ分のクエリログに対し、クエリの属する分野の判定を行った。複数キーワードからなるものや文章であっても一つのクエリとして扱った。分野の判定には、低頻度で検索されるクエリに対しても網羅的にクエリの属する分野の判定が可能なことから、廣嶋ら^[4]の手法を用いた。判定する分野は文献^[4]での実験に準じ、あら

はじめ設定した 27 の分野を用いた．分野判定に用いた 27 分野を表 1 に示す．

各ユーザがそれぞれの分野で投入したクエリと，そのクエリの頻度をまとめ，表示された結果の目視による確認を行った．十分な量の入力回数がある．1 人のユーザが複数種類のクエリを入力している．詳しい人，詳しくない人など，様々な人が入力しうる．クエリ入力回数の分散が大きすぎないといった条件から，スポーツ分野を実験の対象分野として選択した．

表 1. クエリの分野判定に用いた 27 分野

芸能人	車名	政治家	映画
書籍	曲名	地名	ゲーム
番組名	占い	企業家	着メロ
画像	作家	企業名	サイト名
家電	俗語	グルメ	アダルト
健康/医療	スポーツ	アナウンサー	キャラクター
ギャンブル	イベント名	PC 関連	

5. 実験

提案手法の有効性を評価するために，提案手法による推定結果と人手による推定結果の比較実験を行った．被験者の特定分野に関する知識を直接測定することは困難であることから，人手による推定を正解データセットとしている．クエリログからの人手による推定が可能であることを確かめるために，性別，年齢に幅のある複数の評価者に推定をさせ，結果にばらつきがないことを確認する．実験で対象とする分野は，予備実験よりスポーツ分野とした．

5.1. クエリログデータの取得

被験者 51 名に，指定した検索システムを使ってスポーツ分野に関する内容を自由に調べさせ，その際のクエリログを得る．被験者 1 人あたり，80 回以上検索させ，51 人合計で 4411 回の検索を行わせた．検索を行った被験者の内訳を表 1 に示す．

表 2. 検索被験者内訳

年齢	男性(人)	女性(人)	合計(人)
10 代	2	0	2
20 代	17	5	22
30 代	7	5	12
40 代	4	8	12
50 代	0	0	0
60 代	2	1	3
合計	32	19	51

5.2. 人手による知識量の推定

人手による，検索者のスポーツに関する知識量の推定を行った．5.1.節でクエリログを取得した際の被験

者 51 名のスポーツ分野に関する知識量を，別の被験者 5 名にそれぞれ推定してもらった．推定者に対し，検索を行った被験者についてクエリログ以外の情報は与えていない．推定者の内訳を表 3 に示す．

知識量の推定は，推定を行う被験者が，上記の実験で得られた 1 名分のクエリログを眺め，クエリログを残した被験者のスポーツ分野に関する詳しさを推定を，下記に示す図 1 の 5 段階から選択することにより行った．検索者によっては，サッカーや野球などの特定のスポーツに関して多く検索しているものもいるが，全てのクエリは同一分野として評価を行わせた．

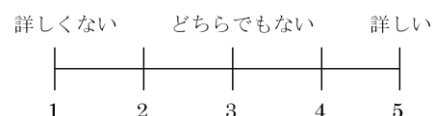


図 1. 詳しさに関する 5 段階評価

推定の際，クエリログ中の知らない語句について調べながら推定をさせた．また 51 人分の推定作業後，推定値順に並べ替えさせ，推定結果に矛盾がないか確認をさせた．5 人の推定結果の一部を表 4 に示す．

5 人の推定結果の分散の平均は 0.76 と十分に小さくなったことから，クエリログから検索者の知識を推定することは可能とし，5 人の推定値の平均を正解データセットとした．

表 3. 推定者の内訳

性別	年齢
女性	30 代
女性	40 代
女性	40 代
男性	30 代
男性	60 代

表 4. 人手による推定結果の一部

検索者番号	推定者 1	推定者 2	推定者 3	推定者 4	推定者 5
1	2	4	3	3	3
2	4	5	3	3	2
3	5	4	5	4	3
4	3	1	3	3	5
5	5	5	3	4	5
6	3	5	5	4	4

5.3. 提案手法による知識量の推定

5.1.節で得られたクエリログを用いて，提案手法による 51 名の被験者のスポーツ分野に関する知識量の推定を行った．提案手法で利用する入力頻度は，国内

の商用検索システムの、2008年6月から2010年1月までの約20カ月分のクエリログから得られたものを利用した。入力頻度のために利用するクエリログ中に、本実験の際に入力されたログは含まれていない。複数のキーワードからなる入力は、1つのキーワードを1クエリとした。また、同一被験者が同じクエリを複数回入力していたものは、1回分のみ利用した。被験者が入力したクエリのうち、ユニークユーザに10回未満しか検索されていないクエリについては利用していない。1人の検索者あたり、平均で約115個のクエリを入力しており、そのうち約91%にあたる約104個を提案手法の計算に利用している。以下の表3に結果の一部を示す。

表 5. 提案手法による推定結果の一部

検索者番号	提案手法による推定値
1	9.86
2	10.7
3	10.8
4	9.02
5	11.9
6	10.2

5.4. 推定結果の比較

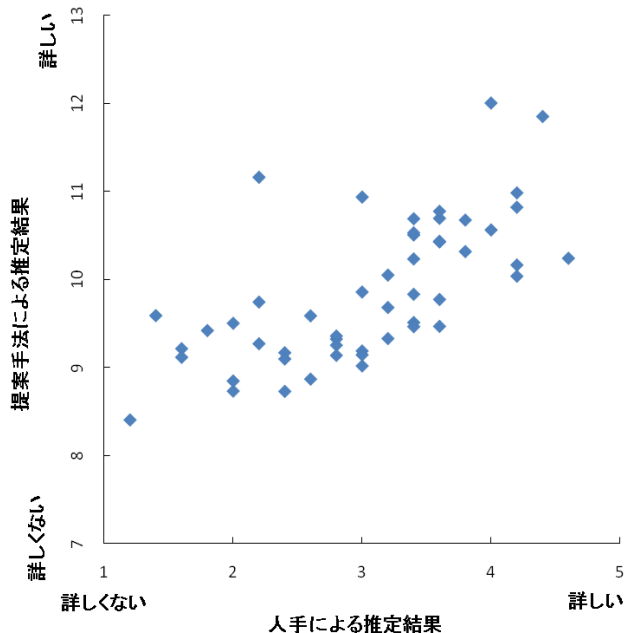


図 2. 人手による推定結果と提案手法による推定結果

人手による推定結果と、手法による推定結果の比較を行った。横軸に人手による推定結果、縦軸に提案手法による推定結果をとったものを以下の図2に示す。人手による推定結果は、5人の推定値の平均をとっている。2つの推定結果の相関係数は0.67であり、提案

法の有効性が確認できた。

2つの推定結果にはいくつか大きな差がある。人手では詳しくないと推定しているが、提案手法では詳しくと推定している場合をみると、検索することにより新たに得られた情報を使って次々と検索を行っている人が判断をしていることが、推定者のコメントからわかる。人手では詳しくと推定しているが、提案手法では詳しくないと推定している場合では、検索者は1回の検索クエリ中に頻度が高いキーワードを複数入れて検索を行っていることが一因であった。

6. まとめ

本研究では、ある分野に関して多くの事柄を知っているということをその分野に詳しいとし、検索システムユーザの分野別の知識量を推定するための手法を提案した。51名の被験者にスポーツ分野に関する事柄について1人約80回程度検索させ、その際のクエリログから各被験者のスポーツ分野に関する知識を、提案手法と人手によってそれぞれ推定を行った。5人の推定値の平均を正解データセットとし、提案手法による推定結果と比較を行った。比較の結果、提案手法と人手による推定値の相関係数は0.67となった。これにより、提案手法が有効であることが確認された。

文 献

- [1] M. Nakatani, A. Jatowt, K. Tanaka: Easiest-First Search, "Towards Comprehension-based Web Search", CIKM'09, ACM, pp. 2057-2060 (2009)
- [2] 西原陽子, 砂山渡, 谷内田正彦, "Web ページの難易度度と学習順序に基づく情報理解支援システム", 電子情報通信学会論文誌 D, Vol. J89-D, No.9, pp.1963-1975 (2006)
- [3] 土方嘉徳, "情報推薦・情報フィルタリングのためのユーザプロファイリング技術", 人工知能学会論文誌, Vol.19, No.3 a, (2004)
- [4] 廣嶋伸章, 戸田浩之, 松浦由美子, 片岡良治, 藤村滋, 森本正志, "概念ベースを利用した Web 検索のクエリタイプ判定", WebDB Forum 2009 (2009)
- [5] 堀幸雄, 中山堯, "ユーザの検索要求に基づいた興味関心の定量的評価", 情報知識学会誌 Vol.16, No.2, pp.33-38 (2006)