L-032

マルチ・クエリー検索に基づくページ検索におけるランク手法

A Ranking Method on Web-page Searching Based on the Multi-query Searching

鈴木 泉†‡ 三上 喜貴† 大里 有生
Izumi Suzuki Yoshiki Mikami Ario Ohsato

1. まえがき

検索をする人間の振る舞いについての研究は、通常の IR システムや、OPAC システムについて 1980 年代から 90 年代にかけて盛んに行われた [1 - 4]。インターネット検索に関しては、検索ログを用いた解析[5] が報告されており、論理演算子の使用は通常の IR システムや OPAC システムに比べ、非常に低いことが確認されている。また、インターネット検索においてユーザーは、検索結果を見るまで、求める情報が何であるかはっきり分かっていない (ill-defined searching) 場合がある[6]。インターネット検索における、ユーザーの知識の不完全さを補完する手法としては、Feedback Searching [8]、Question Answering System [9, 10]、WebFountain [11]、Concept Search [12] などが提案されている。

本研究では、ユーザーが検索しようとしている情報についての知識が十分でなく、common term をクエリーterm として用いなければならない場合を扱う。Single-meaning term つまり、検索しようとする対象で特徴的に使用される term とは異なり、common term をテキスト検索に用いることは適当ではない[7]。しかしその一方、我々は、検索対象の名称などが分からず、対象を説明する言葉を想像し、クエリーを工夫しながら、ページ検索によって求める情報を含むページを見つけることがある。そこで、対象とする検索と問題を第 2 節で明示する。次に、第 3 節で提案するランク手法を提示し、次いで、提案手法と効率を比較するための、検索のヒューマンモデルを提案する。第 5 節では 8 個の実例で検証した結果を示す。

2. 問題の記述

目的は、ページ検索において、ユーザの求める情報を含むページ（answer page と呼ぶ）がより上位に来るように検索結果のランク付けをすることである。対象とするページ検索は以下の通りである。

- 少なくとも 4 個の term をクエリーとして含み、
- 先頭に置かれた term ほど重要と考える。
- 異なる URL であっても、plain text の内容が同じものは、同一のページと考える。

その上で、次の検索（common term による検索）を考える。「ユーザーは、求める情報を言葉（自然文や単語など）で表現するが、single-meaning term をユーザーは知らないものとする。表現された言葉から term を取り出し、term を組み合わせてクエリーを作成する。」

Single-meaning term の例としては、The pitch of the siren of an emergency vehicle becomes lower as it passes by. という記述における、Doppler effect である。

3. 提案手法

ユーザーによって与えられたクエリーは “OR”, “~”, or “NOT といった演算子を含んでも良い。しかし、4 個以上の term $t_1, t_2, \ldots t_n$ については、“AND”演算子のみによって結ばれているとする。

(1) $t_1, t_2, \ldots t_n$ から幾つかの term を取り除くことによって、10 個（ないしは 20 個）のクエリー $q_1, q_2, \ldots, q_{10}$ (or q_{20}) を作成する。その際、先頭の term ほど残されるようにする。

(2) 各クエリーによって検索を実行し、上位 10 ランクに選ばれたページ全体を$\{d_1, d_2, \ldots d_n\}$.とし、$d \in \{d_1, d_2, \ldots d_n\}$. の頻度 f_d とランクの平均値 r_d を求める。

(3) 頻度 f_d について、高頻度の値を丸め、正規化する。

$f'_d = \log(f_d + a) / \log(\mathrm{Max}_d f_d)$, $a = 0.5$ はパラメータ.

その上で、中および低頻度部分に高い評価値を与える。

実験で用いた方法：$f''_d = 3 f'^3_d - 7.2 f'^2_d - 4.8 f'_d$

(4) f''_d および r_d から d のスコア s_d を計算する.

$w_F f''_d + w_R r_d$, w_F および w_R は重み

(5) s_d に従ってページのランクを決定する

4. ヒューマンモデル

ユーザーは、クエリーを $q_1, q_2, \ldots, q_{10}$ の中から無作為に選んで検索するものとする。ただし一度選んだクエリーは二度以上は選ばないものとする。上位 10 ランクに求めるページがなければ、別のクエリーを選択し、検索することを繰り返す。E_i を、i 回目の検索で求めるページを検索する事象とする。p, q をそれぞれ、全クエリー数、求めるページを検索するクエリーの個数とする。E_i が起こる確率は、以下のように帰納的に求められる。

$$P(E_1) = \frac{p}{q}, \quad P(E_2) = P(E_2 \mid \overline{E}_1)P(\overline{E}_1) = \frac{p}{q-1} P(\overline{E}_1)$$

$$P(E_i) = \frac{p}{q-(i-1)} P(\overline{E}_{i-1})$$

次に、求めるページを含む場合のランクの期待値を r 、また、$m = n / 10p$ とする。n は前節で述べた、検索された全ページ数である。i 回目の検索において、検索されたページがこれまで 1 度も現れていない確率を $d(i)$ とし、$d(i)$ は次式によって与えられるものとする。

$$d(i) = -(1-m)/(q-p-1)\, i + 1 + (1-m)/(q-p-1)$$

求めるページを最初に検索するまでにチェックしなければならないページ数 $e(q, p, r, m)$ は以下の式で与えられる。

$$e(q, p, r, m) = \sum_{i=1}^{q-p+1} \left(r P(E_i) + 10 d(i) P(\overline{E}_i) \right)$$

---

† （社）情報処理学会, IPSJ

‡ （社）電子情報通信学会, IEICE

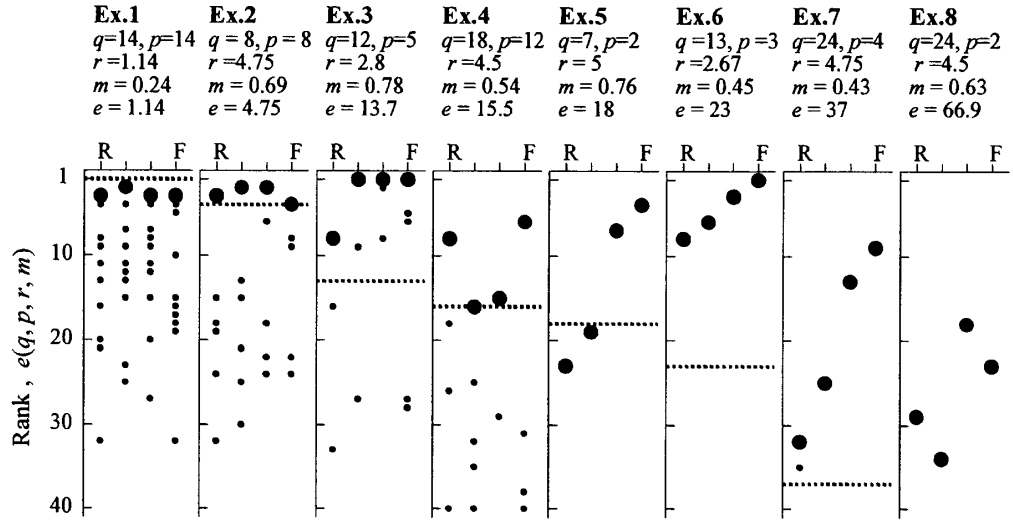

図 1 目的のページが現れるまでに閲覧するページ数の平均 提案手法 (点) とヒューマンモデル (点線) の比較

5. 検証実験

Google を用い、文献[13] を参考に作成した以下の 8 例について実験を行った。Term（下線）の作成、重要度の付加、検索されたページの確認作業は筆者自身が行った。

Ex.1. A musical toy, a flexible, corrugated plastic tube
Ex.2. Cracking a door just a little, the noise level is the same as the door wide open.
Ex.3. The annual (or every year) picking rocks that must be cleared from the garden (or ground) each spring
Ex.4. Fog at the mouth of a champagne or soda bottle just after (or freshly) it's been opened.
Ex.5. The shaft of searchlight beam ends suddenly, not gradually
Ex.6. When I bend down to the ground while listening to an airplane fly by, the pitch of airplane noise increase (or rise)
Ex.7. I have experienced that if I put my head (or ear) close to the ground while listening to an airplane fly by, the pitch of airplane noise increase (or change)
Ex.8 The annual (or every year) crop (or gather) of stones that must be cleared from the garden (or ground) each spring

図 1 は、提案手法によってランク付けした場合の、求めるページのランクを点で示したものである。各事例について、左から順に、重み (w_F, w_R) を (0, 1), (1/3, 2/3), (2/3, 1/3), および (1, 0) とした結果を示している。また、ランクの最も高いページを大きな点で示している。Ex. 1 ないしは Ex. 2 といった容易なケース以外では、特に$(w_F, w_R) = (1, 0)$とした場合に、ヒューマンモデルにおける $e(q, p, r, m)$ よりも大幅に改善されていることが分かる。

6. 考察

提案する手法によってランクが改善されたのは、実験した 8 例に特有の事象ではなく、以下の理由による。

- Ex. 1 ないしは Ex. 2 といった容易なケースでは、求めるページは f_dの広範囲に渡って分布している。
- 逆に、難しいケースでは、求めるページは数が少なく、f_dの中および低頻度部分に分布している。

その一方で、以下の条件を同時に満たすことは起きにくいことから、f_d の高頻度部分のみに求めるページが集中することは起こりにくい。

- ユーザーが作成した term と、answer page で使用されている term が（よく使われる表現などで）類似している。
- Answer page が、Web 上に多数存在しない。

参考文献

[1] Penniman, W.D., 1975. A stochastic process analysis of on-line user behavior, In Information Revolution: Proceedings of the 38th Annual Meeting of the American Society for Information Science, Washington, DC, pp.147-148.
[2] Oldroyd, B.K., 1984. Study of strategies used in online search 5: differences between the experienced and inexperienced searcher, Online Review, Vol. 8 No. 3, pp. 233-244.
[3] Fidel, R., 1984. Online searching styles: a case-study-based model of searching behavior, Journal of the American Society for Information Science, Vol. 35, pp: 211-221
[4] Hsieh-Yee, I., 1993. Effects of search experience and subject knowledge on the search tactics of novice and experienced searchers. Journal of the American Society for Information Science, Vol. 44, No. 3, pp.161-174.
[5] Jansen, B. J. et al, 2000. real users, and real needs: a study and analysis of user queries on the web, Information Processing and Management, Vol. 36, pp.207-227.
[6] Saito, M. and Ohmura, K., 1998. A Cognitive Model for Searching for Ill-defined Targets on the Web – The Relationship between Search Strategies and User Satisfaction –, Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Melbourne, pp.24-28
[7] Fidel, R., 1991. Searchers' selection of search keys: I. The selection routine, Journal of the American Society for Information Science, Vol. 42, No. 7, pp. 490-500.
[8] Anick, P., 2003. Using terminological feedback for web search refinement: a log-based study, Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval, Toronto, pp. 88-95
[9] Waltz, D. L., 1978. An English language question answering system for a large relational database, Communications of the ACM archive, Vol. 21, No. 7, pp. 526-539.
[10] Radev, D. et al, 2005. Probabilistic question answering on the Web, Journal of the American Society for Information Science, Vol. 56, No. 6, pp. 571-583.
[11] Gruhl, D. et al, 2004. How to build a WebFountain: An architecture for very large-scale text analytics, IBM Systems Journal, Vol. 43, No. 1, pp. 64-77
[12] Yan, N. and Khazanchi, D., 2007. Concept Level Web Search Via Semantic Clustering, Computational Science - ICCS 2007, Proceedings of the 7th International Conference, Beijing, Part III, pp. 806-812
[13] Walker, J., 1977. The Flying Circus of Physics with Answers. John Willey & Sons, Inc., USA.