

科学研究費助成事業 研究成果報告書

平成 27 年 4 月 16 日現在

機関番号：14301

研究種目：基盤研究(A)

研究期間：2012～2014

課題番号：24240013

研究課題名(和文) ウェブ検索の意図検出と多元的検索意図指標にもとづく検索方式の研究

研究課題名(英文) Web Search Intent Detection and Query Processing Based on Various Search Intents

研究代表者

田中 克己 (TANAKA, Katsumi)

京都大学・情報学研究科・教授

研究者番号：00127375

交付決定額(研究期間全体)：(直接経費) 35,700,000円

研究成果の概要(和文)：ウェブ検索エンジンは、現代社会において既に、ネット上の情報を検索するためのツールとして欠くべからざるものとなっており、世界中の人々が多くこれを用いてネット情報の検索を行っている。現状のウェブ検索エンジンは、入力したキーワード語を含むウェブページを検索するものであり、多様な検索意図をもつ検索要求には答えられないという欠点がある。本研究では、検索意図として、典型性・稀少性、斬新性・意外性、客観性・主観性、具体性・抽象性、理解容易性、信頼性等に注目し、このような検索意図をもつ検索を実現するために必要となる検索方式を開発した。

研究成果の概要(英文)：Web search engines are now indispensable IT tools for people in our society to search for information on the Internet. Conventional search engines, however, just facilitate to find Web pages that contain keywords specified by users. On the other hand, people usually have a variety of search intents, for example, "they wish to find a typical/novel/objective/concrete/comprehensive/credible information on certain topics". In this research, we first investigate search intents, and explored search algorithms suitable for these various search intents.

研究分野：情報検索

キーワード：ウェブ 情報検索 意図検出

1. 研究開始当初の背景

ウェブ (World Wide Web、以下 Web と略記) 上で発信される情報は、社会、経済、政治、文化、学術、日常生活など、社会のあらゆる方面における重要な情報源となっており、Web 検索エンジンは、Web データにアクセスするために欠くことのできない基盤技術となっている。

従来の Web 検索エンジンや検索サービスでは、利用者の検索質問は検索語で表現され、入力された検索語を含む Web 文書が結果として返されるが、検索結果が大量であり、そのため何らかの評価指標によって検索結果を順位付け (ランキング) するための技術が不可欠となっている。

一方、利用者の検索要求や検索意図は極めて多様である。Google 等で採用されている、クエリと文書内容の類似性や、人気度 (多くの他文書からの参照リンクを受けている度合いを人気度ととらえる) にもとづくランキングだけでは、利用者の多様な検索意図に応えることが難しい。実際、現行の Web 検索エンジンでは、例えば、京都観光という検索質問に関して、「典型的な文書」、「重要な話題を網羅している文書」、「理解しやすい文書」、「具体的な記述を含む文書」を求めるといった検索要求に応えるのは困難である。

2. 研究の目的

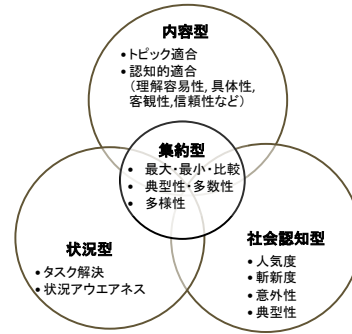
本研究の目的は、ウェブ情報検索における多面的な検索意図指標の確立、各検索意図指標にもとづく検索方式、および、検索質問 (クエリ) やユーザインタラクションに基づく検索意図の自動検出 (Search-Intent Detection) 方式を開発することである。検索意図指標として、従来から用いられているクエリと文書間の類似度や人気度 (popularity) に加えて、典型性 (typicality)、多様性 (diversity)、理解容易性 (comprehensibility)、具体性 (concreteness) など、多様な検索意図指標とその検出方法を開発するとともに、これらに応じた効果的・効率的な検索が可能となる情報検索モデルと検索方式を開発する。

3. 研究の方法

検索意図指標 (正解判定指標) として、文書の、クエリ類似性、人気度、典型性、多様性、理解容易性、具体性など、多様な指標概念の確立を目標に、Web 検索エンジンの利用者の検索意図の調査研究、これらの多様な検索意図指標に応じた検索・ランキング方式の開発、および、人・システムの相互作用にもとづき検索者の検索意図推定的方式研究を行う。

4. 研究成果

(1) 検索意図指標の確立



本研究では、Saracevic[1]の研究をもとに、情報検索の「検索要求(意図)」を、上図のように再分類した。

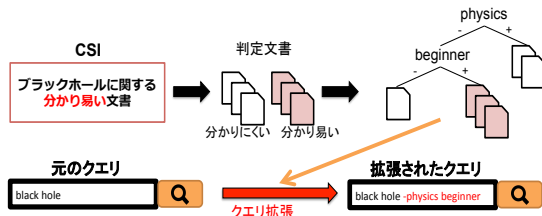
ここで、内容型検索要求とは、文書の内容に関する検索要求であり、クエリに書かれた話題 (キーワード) と類似の話題の文書を検索したいという検索要求 (トピック適合) と、指定した認知的特性を備えている文書を検索したいという検索要求 (認知的適合) である。前者は、従来の情報検索における狭義の適合性にあたり、後者は、「理解しやすい」、「客観的」、「信頼できる」等といった認知的な特性を備えている文書を検索するという検索要求である。

状況型検索要求とは、問題解決を目的とするような検索であり、検索クエリはそのユーザの達成目標や状況を記述したものとなる。例えば、「パソコンに習熟する」というタスクを達成するための方法を検索する場合や、「子供の誕生日にふさわしいレシピ」を検索するような状況アウェアな検索が相当する。社会認知型検索要求とは、文書 (情報) に対する社会的な認知性に基づいて検索する場合が相当する。例えば、「人気のある観光地」、「典型的な観光地」、「意外な観光地」等の検索である。

集約型検索要求とは、上記の3種の検索要求の検索を行って得た結果をさらに「集約」した情報を最終的に求めるような検索がこれにあたる。例えば、「最も人気のある観光地」、「信頼できるホテルAの評価」等の検索がこれにあたる。

(2) 検索意図指標に応じた検索方式の開発 (2-1) 内容型検索要求

理解容易性 (分かり易さ) 等の認知的適合性に基づく検索を実現するために、クエリで用いるキーワードを (決定木を用いた) 機械学習により発見する基本的な手法を開発し、WWW2014、ACM SIGIR2014 等のトップコンファレンスで発表した。この手法は、認知的に適合する文書としない文書を訓練データとして機械学習 (決定木) を行い、得られた決定木を用いて元のクエリを拡張するものである。下図にその適用例を示す。



種々の認知的な検索意図に応じて、本手法で発見出来たキーワード語を以下の表で示す。

認知的検索意図	トピック独立	%	トピック依存	%
網羅的	details	13.4	type	1.5
	wiki	6.7		
分かり易さ	easy-to-understand	6.6	domestic medicine	0.6
	concrete example	3.0	high school biology	0.6
客観的	wiki	8.8	data	1.8
	details	4.4	scientific data	0.9
主観的	blog	10.2	impression	20.3
	opinion	2.3	usability	3.9
具体的	concrete example	15.3	real experience	0.8
	example	8.9	real world	0.8
抽象的	summary	8.3		
	wiki	7.1		

(2-2) 状況型検索要求

問題解決を目的とする検索では、ユーザの検索要求はそのユーザの達成目標を記述したもの（例えば、「パソコンに習熟する方法が知りたい」、「ギターを売却したい」）や、そのユーザの状況を記述したものになる。また、その検索解も問題解決を行うための方法に関する情報（例えば、「パソコンの基本操作を学ぶ」、「オフィスソフトの使い方を学ぶ」や「中古ギターを購入する業者を探す」など）となる。すなわち、「〇〇を△△する」という形のタスクそのものがクエリとなり、検索解もこのタスクを達成するためのタスク群が解となる。このような状況型検索をタスク検索と呼ぶと、タスク検索においては、元の検索要求タスクのパラフレーズ（例えば、「パソコン、習熟」という元の検索要求に対して、「PC、マスター」等のパラフレーズ表現の検索）が必要となるが、さらに、それに加えて、入力タスクを遂行するのに必要となる動作名やオブジェクト名による検索語の置き換えが重要となる。これは、元のタスクの動作・動詞と含意 (entailment) 関係にある動作・動詞（「習熟」に対して「学習」など）や、元のタスクの主体客体を逆転させた動作・動詞を発見（「売却」に対して「購入」など）し、このような動作名・動詞で検索質問を変換することに対応する。

本研究では、このようなタスク検索を実現するために、与えられたクエリから適切なクエリ語群をブートストラップ検索で発見する方式を開発した。ブートストラップ検索では、冗長なクエリ語 (redundant terms) や意味的に不適切なクエリ語 (semantic drift terms) までもが検索されてしまうという問題が起こるため、これに対して、検索された語群の多様性分析を行うことで解決した。

(2-3) 社会認知型検索要求

語の認知度と語間の関係の非典型度を用いて Wikipedia から「意外」な情報を発見する検索方式を開発した。

(2-4) 集約型検索要求

検索結果のページの典型性計算については、認知心理学分野の Barsalou の提唱する 3 種の典型性 [2] (CT 典型性: 類似したオブジェクトが多いオブジェクトほど典型的、I 典型性: あるカテゴリが究極的に満たすべき状態を満たしているオブジェクトほど典型的)、FOI 典型性: 多くの人間が認知しているオブジェクトほど典型的) に対応する計算方式を開発した。

検索結果のページの信頼度を算出するために、その情報源・情報発信者・情報内容・情報の観点がどの程度多様であるかを計算する方式を開発した。多様性計算は、直観的には、「多様な情報発信者・情報源から言及・参照されている情報は信頼できる」という仮説に基づき、収集されたデータの発信者多様性などを計算するものである。例えば、ユーザ個人個人が発信する（宿泊施設、グルメ、観光地、映画等の）レビュー情報は、現在は ISP 側で頻度別集計してユーザに提示されているが、レビューした発信者のプロフィールやレビューの観点の多様性は無視されている。本研究では、情報発信者・情報源の多様性を、相違性 (disparity 多様性)、均衡性 (balance 多様性)、多種性 (variety 多様性) 等の種々の多様性尺度に基づいて分析する手法を開発した。

(3) 検索行動分析による検索意図抽出

質問応答サイトの情報や Web 広告情報から検索意図に関わる情報を抽出する方式、検索結果や該当ページに対するユーザの視線情報を分析して検索意図を推定する方式、ユーザのクエリ履歴を分析してクエリの修正を予測する方式、ユーザの検索スキルと事前知識に着目して検索行動とユーザの検索タスク満足度の関係性を分析する方式、検索システム側からユーザに適切な質問を行う事でユーザの検索目標を明確化する方式などを開発した。



ユーザの視線情報を分析して検索意図を推定する画面を上図に示す。ここでは、ユーザの検索クエリはイタリアの観光地 capri であ

るが、ユーザの視線は、主に宿泊施設（ホテル）の部分に集中していることから、ユーザの検索意図を、イタリアの観光地の宿泊施設であると推定している。

<引用文献>

[1] T. Saracevic: Relevance Reconsidered, Information science: Integration in perspectives, Proc. 2nd CoLIS, pp. 201-218, Oct. 1996.

[2] L. W. Barsalou: Ideals, central tendency, and frequency of instantiation as determinants of graded structure in categories, J. Experimental Psychology, pp. 629-654 (1985).

5. 主な発表論文等

[雑誌論文] (計 11 件)

- [1] 門脇 拓也, 山肩 洋子, 森 信介, 田中 克己: 誕生・使用事由によるレシピ検索～生い立ちレシピサーチ～, 日本データベース学会和文論文誌 13(1), 78-85, 2014 年 10 月
- [2] 梅本和俊, 山本岳洋, 田中克己: 検索専門性と事前知識に着目した検索行動とタスク満足度の関係性分析, 情報処理学会論文誌 (トランザクション) データベース (TOD64), 7(4), 13-28, 2014 年 12 月.
- [3] 佃洗撰, 大島裕明, 山本光穂, 岩崎弘利, 田中克己: 語の認知度と語間の関係の非典型度に基づく Wikipedia からの意外な情報の発見, 情報処理学会論文誌 (トランザクション) データベース (TOD61), 7(1), 1-17, 2014 年 3 月
- [4] Makoto P. Kato, Tetsuya Sakai, Katsumi Tanaka: When do people use query suggestion? A query suggestion log analysis. Inf. Retr. 16(6): 725-746 (December 2013)
- [5] 梅本和俊, 中村聡史, 山本岳洋, 田中克己: Web 検索時の行動情報を用いたクエリ修正タイプの予測, 情報処理学会論文誌: データベース (TOD58), 6(3), 132-147, 2013 年 6 月.
- [6] 梅本和俊, 山本岳洋, 中村聡史, 田中克己: 視線情報からの注目語抽出に基づく検索意図のリアルタイム推定, 情報処理学会論文誌: データベース (TOD58), 6(3), 120-131, 2013 年 6 月.
- [7] 山本祐輔, 田中克己: 反証センテンスの提示による信憑性指向のウェブ検索支援, 情報処理学会論文誌: データベース, 6(2), 42-50, 2013 年 3 月
- [8] 佃洗撰, 大島裕明, 山本光穂, 岩崎弘利, 田中克己: 語の認知度と同位語間の関係に基づく意外な情報の発見, 日本データベース学会論文誌, 11(3), 21-26,

2013 年 2 月

[学会発表] (計 42 件)

- [1] Shuya Ochiai, Makoto P. Kato, Katsumi Tanaka: Re-call and Re-cognition in Episode Re-retrieval: A User Study on News Re-finding a Fortnight Later. Proc. 23rd ACM International Conference on Conference on Information and Knowledge Management (CIKM2014), pp. 579-588, Shanghai, China, November 3-7, 2014.
- [2] Yoshiyuki Shoji, Makoto P. Kato, Katsumi Tanaka: Can Diversity Improve Credibility of User Review Data? , Proc. 6th International Conference on Social Informatics (SocInfo2014), Springer, Lecture Notes in Computer Science, Vol. 8851, pp. 244-258, Barcelona, Spain, Nov. 11-13, 2014.
- [3] Yoshinori Kitaguchi, Hiroaki Ohshima, Katsumi Tanaka: Query Formulation for Action Search by Bootstrapping, Proc. 16th International Conference on Asia-Pacific Digital Libraries (ICADL 2014), Springer, Lecture Notes in Computer Science, Vol. 8839, pp. 417-422, Chiang Mai, Thailand, Nov. 5-7, 2014.
- [4] Takuya Kadowaki, Shinsuke Mori, Yoko Yamakata, Katsumi Tanaka: Recipe search for blog-type recipe articles based on a user's situation, Proc. 2014 ACM Conference on Ubiquitous Computing, UbiComp '14 Adjunct, pp. 497-506, Seattle, WA, USA, Sept. 13 - 17, 2014.
- [5] Kosetsu Tsukuda, Hiroaki Ohshima, Katsumi Tanaka: Ranking of Coordinate Terms and Hypernyms Using a Hypernym-Hyponym Dictionary, Proc. IEEE/WIC/ACM International Joint Conferences on Web Intelligence (WI) and Intelligent Agent Technologies (IAT), (Vol. 2), pp. 15-21, Warsaw, Poland, Aug. 11-14, 2014.
- [6] Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka: Investigating Users' Query Formulations for Cognitive Search Intents, Proc. 37th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR2014), pp. 577-586, Gold Coast, Australia, July 6-11, 2014.
- [7] Kosetsu Tsukuda, Tetsuya Sakai, Zhicheng Dou, Katsumi Tanaka: Estimating Intent Types for Search Result Diversification, Proc. 9th

- Asia Information Retrieval Societies Conference (AIRS 2013), Singapore, Springer, Lecture Notes in Computer Science, Vol. 8281, pp.25-37, Dec. 09-11, 2013.
- [8] Yoshiyuki Shoji, Katsumi Tanaka: Diversity-Based HITS: Web Page Ranking by Referrer and Referral Diversity, Proc. of the 5th International Conference on Social Informatics (SocInfo2013), Kyoto, Japan, Springer, Lecture Notes in Computer Science, ISBN 978-3-319-03259-7, pp.377-390, November 25-27, 2013.
- [9] Katsumi Tanaka: Can We Predict User Intents from Queries? - Intent Discovery for Web Search (Keynote Talk), The 18th International Conference on Database Systems for Advanced Applications, April 22-25, 2013, Wuhan, China.
- [10] Shinya Tanaka, Adam Jatowt, Makoto P. Kato, Katsumi Tanaka: Estimating Content Concreteness for Finding Comprehensible Documents, Proc. 6th ACM International Conference on Web Search and Data Mining (WSDM2013), Rome, Italy, pp. 475-484, Feb. 4-8, 2013.
- [11] Soungwoong Yoon, Adam Jatowt, Katsumi Tanaka: Search Intent Discovery by Structurization of Community QA Contents, Proc. 13th International Conference on Web Information Systems Engineering (WISE2012), Paphos, Cyprus, Lecture Notes in Computer Science, Vol.7651, pp.712-718, Nov. 28-30, 2012.
- [12] Takehiro Yamamoto, Tetsuya Sakai, Mayu Iwata, Chen Yu, Ji-Rong Wen, Katsumi Tanaka: The wisdom of advertisers: mining subgoals via query clustering, Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), ACM Press, Maui, Hawaii, pp.505-514, Oct. 29 - Nov. 02, 2012.
- [13] Adam Jatowt and Katsumi Tanaka: Is Wikipedia Too Difficult? Comparative Analysis of Readability of Wikipedia, Simple Wikipedia and Britannica, Proc. 21st ACM International Conference on Information and Knowledge Management (CIKM 2012), ACM Press, Maui, Hawaii, pp. 2607-2610, Oct. 29 - Nov. 02, 2012.
- [14] Makoto P. Kato, Hiroaki Ohshima, Katsumi Tanaka: Content-based Retrieval for Heterogeneous Domains: Domain Adaptation by Relative Aggregation Points, Proc. ACM 35th Annual SIGIR Conf (SIGIR2012), Portland, USA, pp.811-820, Aug. 12-16 2012.
- [図書] (計2件)
- [1] Adam Jatowt, Ee-Peng Lim, Ying Ding, Asako Miura, Taro Tezuka, Gaël Dias, Katsumi Tanaka, Andrew J. Flanagan, Bing Tian Dai (Eds.): Social Informatics - 5th International Conference, SocInfo 2013, Kyoto, Japan, November 25-27, 2013, Proceedings. Lecture Notes in Computer Science 8238, Springer 2013.
- [2] Dietmar Jannach (著), Markus Zanker (著), Alexander Felfernig (著), Gerhard Friedrich (著), 田中克己, 角谷和俊 (監訳): 情報推薦システム入門-理論と実践-, 359 ページ, 共立出版, 2012年6月22日. (ISBN-10: 4320122968, ISBN-13: 978-4320122963)
- [その他]
- ホームページ等
 京都大学情報学研究科田中研究室業績
<http://www.dl.kuis.kyoto-u.ac.jp/wordpress/publication>
6. 研究組織
- (1) 研究代表者
田中 克己 (Tanaka, Katsumi)
 京都大学・情報学研究科・教授
 研究者番号: 0012375
- (2) 研究分担者
アダム ヤトフト (Jatowt, Adam)
 京都大学・情報学研究科・特定准教授
 研究者番号: 00415861
- 中村 聡史 (Nakamura, Satoshi)
 京都大学・情報学研究科・特定准教授
 (現、明治大学・総合数理学部・准教授)
 研究者番号: 00415861
- 大島 裕明 (Ohshima, Hiroaki)
 京都大学・情報学研究科・特定准教授
 研究者番号: 50415858
- 加藤 誠 (Kato, Makoto)
 京都大学・情報学研究科・特定助教
 研究者番号: 00646911
- 山本 岳洋 (Yamamoto, Takehiro)
 京都大学・情報学研究科・助教
 研究者番号: 70717636