

NIIとYahoo! JAPANが検索技術研究のために新たな提携 「Yahoo!検索」の検索クエリデータをNIIのワークショップに無償提供

国立情報学研究所

ヤフー株式会社

ヤフー株式会社（以下Yahoo! JAPAN、代表取締役社長：宮坂 学、東京都港区）と大学共同利用機関法人 情報・システム研究機構 国立情報学研究所（以下NII、所長：喜連川 優、東京都千代田区）は、情報学研究の一層の推進に寄与するために、Yahoo! JAPANよりNIIに「Yahoo!検索」の検索クエリデータを無償提供することになりました。本データは、情報アクセス技術の評価ワークショップ「NTCIR（エンティサイル）」（*1）の参加者向けに提供され、同ワークショップに参加する研究グループは無償で活用できます。

Yahoo! JAPANが提供するデータは、2009年7月から2013年6月の期間に「Yahoo!検索」で検索された全クエリ（ユーザーが検索時に入力した単語やフレーズ）の中から、現在進行しているNTCIRの第12サイクル（NTCIR-12）で設定された研究課題に対する関連度の高いクエリを抽出したものです。本データで使用しているのは異なる数十人以上のユーザーが検索に用いたクエリに限られ、「Yahoo!検索」ユーザー個人の操作履歴や識別子、属性といった個人情報は一切含んでいません。

NTCIRの実行委員会で共同委員長を務めるNII情報社会相関研究系教授の神門 典子は、今回NTCIRに「Yahoo!検索」のデータが提供される意義について、「ユーザーが入力した検索クエリの背後にある意図を推定して多様な検索結果を提示するといった、現在のウェブサーチエンジンの検索有用性を左右する主要技術の研究には、ユーザーの検索クエリデータが不可欠です。このため、検索サービスを提供している企業以外はなかなか研究できない状況でした。今回のデータ提供は、世界でも極めて珍しく、検索サービス提供企業以外でもこうした研究が可能になりました」と話しています。

NTCIR-12ではこれらの貴重なデータを、検索意図のマイニングやモバイル用に様々な検索意図を満たすように多様化した検索結果を二段階で要約する技術の研究で活用することになっています。Yahoo! JAPAN提供のデータを使用するタスクへの参加は、7月31日まで受け付けています。

検索クエリデータは、ユーザーの検索意図の理解、レコメンデーション、広告関連技術など、さまざまな検索から派生する技術研究において活用できます。さらに、本データを活用した研究により、近未来のロボットや家電、ウェアラブルなどに組み込まれるであろう「人間と対話するコンピュータ技術」を大きく発展させられる可能性があります。Yahoo! JAPANはこれまでもNIIに対して、「Yahoo!知恵袋」で解決済みとなった質問と回答の投稿データなどを2007年3月と2009年6月の2回にわたって提供しています。（*2）

NIIでは、Yahoo! JAPANから提供されたデータを活用したNTCIR-12をはじめとした研究活動を通じて、より豊かな情報アクセス技術の実現と情報学分野の未来価値の創成に取り組んでいます。また、2017年に東京で開催される情報検索のトップカンファレンス「ACM SIGIR」(*3)に向けて、今回のYahoo! JAPANからのデータ提供のように、さまざまな形で産学が協力し、日本における情報検索研究を一層促進できるように努めてまいります。

(*1) 「NTCIR」：正式名称は「情報検索システム評価用テストコレクション構築プロジェクト (NII Testbeds and Community for Information access Research)」。検索システムの性能評価という切り口から、情報検索やテキスト要約、テキストマイニング、意見や動向の分析など膨大なデータの中から求める情報にアクセスするためのテキスト処理技術の研究に取り組むNIIのプロジェクト。国内外の研究機関や企業に所属する研究者が企画運営にあたる。サイクルごとにタスク（研究部門）を設定、参加する研究グループは共通の基盤上でそれぞれの研究を進めるとともに、お互いに検証や評価をすることで研究を集中的に推進する。1997年に立ち上がり、1998年11月に始まった第1サイクル (NTCIR-1) を皮切りに、概ね約1年半のサイクルでワークショップを開催。各サイクルの最後には「NTCIR カンファレンス」を国際会議として開催している。NTCIR-11 (2013年9月～2014年12月) には日本を含む20カ国・地域から計132研究グループが参加。

(*2) 「『Yahoo!知恵袋』データのNIIへの提供」：第1版 (2007年3月提供) は対象期間が2004年4月～2005年10月、質問数約300万、回答数約1300万、第2版 (2009年6月提供) は対象期間が2004年4月～2009年4月、質問数約1600万、回答数約5000万。

(*3) 「ACM SIGIR」：米国をベースとする計算機科学の国際学会「ACM」 (Association for Computing Machinery) の情報検索 (Information Retrieval) 分野に関する分科会 (Special Interest Group=SIG) 。