NEWS RELEASE

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 National Institute of Informatics

平成 28 年（2016 年）2 月 19 日

検索クエリのパターン抽出の効率的手法を開発
Web 検索ユーザーの意図の推測根拠をより簡単・高速・的確に

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（NII、所長：喜連川 優、東京都千代田区）のビッグデータ数理国際研究センター 特任研究員 小西 卓哉、同 特任研究員 大輪 拓也（現株式会社富士通研究所）、同 特任助教 林 浩平と、ヤフー株式会社 Yahoo! JAPAN 研究所の藤田 澄男上席研究員、奈良先端科学技術大学院大学 情報科学研究科の池田 和司 教授は、「JST ERATO 河原林巨大グラフプロジェクト」(*1) の研究の一環として、ヤフー株式会社（東京都港区）と共同で、Web 検索で使用される検索クエリ（問い合わせ）からユーザーの意図を推測するためのパターン（傾向）を抽出する新たな手法を開発しました。

この結果は、現地時間の 2 月 22 日から米国・サンフランシスコで開催される Web 検索とデータマイニングに関する最重要の国際会議「WSDM2016」(*2) で発表されます。

検索クエリに込められたユーザー意図の推測は、Web検索システムの機能を考える際の重要な検討課題になっています。ユーザーの興味を把握できれば、それに従って検索結果を並び替えることで、Web検索の性能を改善できます。また、ユーザーの興味や嗜好に基づくターゲット広告などのマーケティングも、より高精度かつ動的に行うことが可能となります。

本共同研究では、ユーザー意図推測の根拠となる検索クエリのパターンを、機械学習の技術を応用して発見する新たな手法を開発しました。この手法を使うと、ヒトの知識に頼ることなく完全に自動的にパターンの抽出を行い、対象とする検索クエリデータが大規模になっても、ユーザー意図を高精度に推測できる可能性を示しました。

National Institute of Informatics

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
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp

Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

【背景と既存の技術】

検索クエリからユーザー意図を推測するために、その中に含まれる単語の組み合わせパターンを利用することができます。例えば、「箱根 旅館 安価」と「京都 ホテル 駅チカ」という二つのクエリは、ともに『地名』、『旅行』、『条件』というトピック（題目）が組み合わせられた『地名 旅行 条件』というパターンだと考えることができます。こうした検索クエリを入力したユーザーの意図は宿泊施設を探すことだと推測できます。

このパターン抽出については、既にいくつかの方法が開発されています。ある研究は機械学習における確率モデルを用いた抽出法を提案しています。この手法では、検索クエリに含まれる単語とそれに対応するトピックの組み合わせを全て調べることでパターンを抽出します。しかし、トピック数が多いと、この組み合わせの数が爆発的に増加します。一般に検索クエリは様々なジャンルの単語を含んでいるため、トピック数の設定値を大きくする必要があり、トピックの組み合わせの計算に時間がかかることが問題となります。この研究では、検索クエリのジャンルを予め人手を含む前処理で限定することで処理速度の問題を回避していましたが、今度は、前処理を行うことによるコストの増加やジャンルの限定に伴う性能の劣化など他の問題が発生してしまいます。

また、同じく確率モデルであるトピックモデルをパターン抽出に応用する方法も考えられています。トピックモデルはテキストデータから各単語のトピックを推測する手法であり、比較的トピック数が大きい設定でも使用できます。ただ、トピックモデルはニュース記事のように多くの単語を含むテキストデータを想定していますが、検索クエリの場合、一つひとつは非常に短いテキストデータです。このため、通常のトピックモデルを応用するとパターンを構成するトピックとして的確なものを推測できないという、精度の問題がおこります。近年、短いテキストデータのためのトピックモデルも提案されていますが、このモデルはトピックの組み合わせを考慮しないため、こちらも的確なトピックが推測できないことが問題になります。

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
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp
Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

【今回開発した手法の特徴】

今回の共同研究では、トピックの組み合わせを単語のペアに限定する新たなトピックモデルを考案しました。先に例に挙げた検索クエリ「箱根 旅館 安価」の場合、３単語それぞれに対応するトピックの組み合わせを調べなくても、「箱根 旅館」のような単語のペアに対応するトピックのペアだけ調べれば、どのトピックが適切か一定程度の評価をすることができます。このアイデアに基づき、検索クエリ内の全ての単語ペア（「箱根 旅館」「旅館 安価」「箱根 安価」）を抽出することで、頻出するパターンの確率が計算できます。この手法では、人手を含む前処理を必要としないため、パターンの抽出を自動化できます。また、トピックのペアだけを調べるため、計算時間を抑えることができ、処理を高速化することもできます。さらに、トピックのペアを通じてトピックの組み合わせを考慮できるため、従来のトピックモデルと比べて、より的確なトピックとパターンの抽出が可能になります。

この手法の有効性を確認するため、本共同研究では、ヤフー株式会社に蓄積された検索クエリデータを用いて、ジャンルを限定しない一般的な検索クエリデータからトピックとパターンを抽出し、既存のトピックモデルによる抽出結果と精度を比較する実験を行いました。その結果、抽出されたトピックやパターンを人手で評価する問題などで、この手法が既存のトピックモデルよりも高い精度を示すことを確認しました。

今回開発した手法の特徴は、実際の検索クエリデータのみから自動的にトピックを学習するところです。これは、人為的な方法や辞書などの知識によってトピックやパターンを構築するよりも有効な場合があります。例えば、「箱根」「京都」と「横浜市」「世田谷区」はいずれも地名を表す単語ですが、前者が旅行先として挙げられることが多いのに対して、後者は現地に住んでいる住民が地域サービスを調べる際に使用されることが多い単語です。それぞれ「旅行」と「地域サービス」という異なる意図で使われることが多いため、広告などへの応用を考えた際は、「箱根」「京都」と「横浜市」「世田谷区」を別々のトピックとして抽出できることが望まれます。しかし、人為的な手法や辞書などの知識を使うと、これら４単語は全て「地名」という同一のトピックとして扱われてしまう可能性があります。これに対し

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
総務部企画課 広報チーム
〒101-8430 千代田区一ツ橋 2-1-2
直通：03-4212-2164 FAX：03-4212-2150
E-Mail: media@nii.ac.jp
Web: http://www.nii.ac.jp
Twitter: @jouhouken
facebook: https://www.facebook.com/jouhouken

て、今回開発した手法は、実際の検索クエリデータの偏りから自動的に分類を行うため、より細かいトピックを抽出できます。このため、先の例のように同じ「地名」の単語でも、使われた意図にそって異なるトピックとして抽出してパターンを推測することを可能になりました。

以上

〈メディアの皆様からのお問い合わせ先〉

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

総務部企画課 広報チーム（担当：美土路昭一）

TEL：03-4212-2164　E-Mail：media@nii.ac.jp

（*1）「JST ERATO 河原林巨大グラフプロジェクト」：国立研究開発法人 科学技術振興機構（JST）の戦略的創造研究推進事業・総括実施型研究（Exploratory Research for Advanced Technology=ERATO）に採択された研究プロジェクト。NII 情報学プリンシプル研究系教授、河原林健一が研究総括を務める。

（*2）「WSDM2016」：米国をベースとする計算機科学の国際学会「Association for Computing Machinery」（ACM）が開催する「The 9th ACM International Conference on Web Search and Data Mining」。会期は 2 月 22 日～25 日。