プレスリリース 2016.11.24

Yahoo! JAPAN、ビッグデータ分析の高速化を実現する世界最速クラスの高次元データ検索技術「NGT」を特許実施権無償提供の上、オープンソースソフトウェアとして公開開始

～ 主流の既存類似技術の12倍を超える検索速度を実現
ディープラーニングを用いたコンテンツ配信精度のさらなる向上
や
AI（人工知能）分析に適した活用しやすいビッグデータ生成に貢献 ～

ヤフー株式会社は、ビッグデータ分析領域の最先端技術として、高次元データの高速検索技術「NGT（Neighborhood Graph and Tree for Indexing）」を開発し、本日よりオープンソースソフトウェア（以下、OSS）として公開します。なお、同技術に関係する特許の実施権を無償で提供します。

■「NGT」概要

「NGT（Neighborhood Graph and Tree for Indexing）」は、テキストや画像、商品・ユーザーデータなど、複数の特徴を持つデータ（高次元データ）を、大量のデータベースの中から、高速に検索・特定できる技術で、AI（人工知能）やIoTの台頭により、ますます巨大化の一途をたどるビッグデータの分析の高速化を実現します。

本技術は、言語データや画像特徴データ、いずれにおいても、技術の最先端である学術領域における既存の類似技術の中でも主流の技術と比べて、12倍以上も高速に検索できることが確認されています。（※）

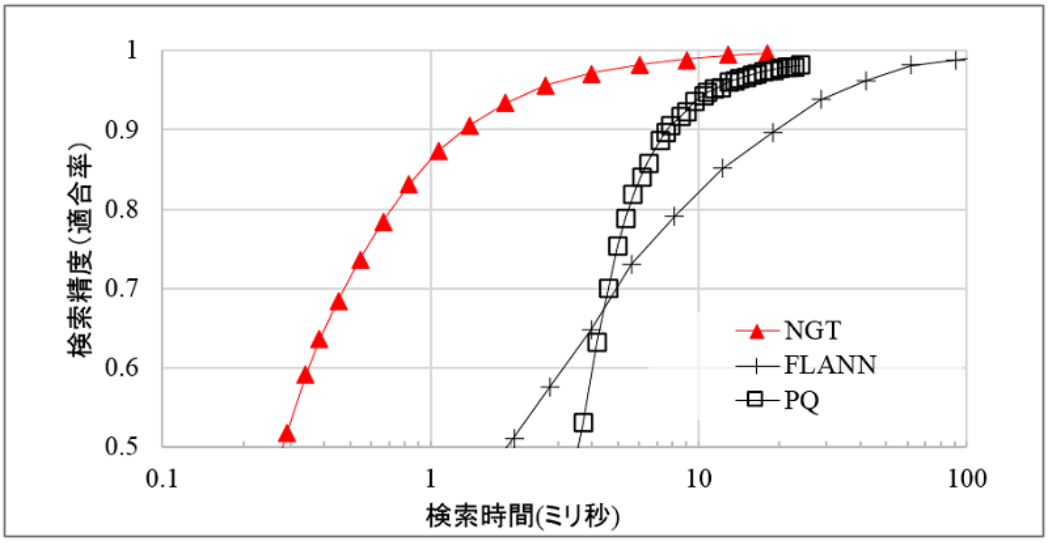
特に、言語データにおける最新研究結果は、AI（人工知能）における重要領域の一つである“自然言語処理”分野において最高峰と言われる国際会議「ACL 2016 (54th Annual Meeting of the Association for Computational Linguistics)」（2016年8月開催）でも論文として採択され、その有用性が高く評価されました。

なお、商用不可の研究用途に限定した形で、既に2015年9月より「Yahoo! JAPAN研究所」サイト上にて公開していましたが、このたび、社内外問わず多くのデータサイエンティストとともに本技術を発展させていくため、本日より特許実施権を無償提供の上OSSとして、広くエンジニアに活用されているソフトウェア共有サイト「GitHub」上に公開します。

※【言語データ研究のエビデンス】
検索精度（適合率）90%として、200万件の言語データを対象に検索を行った場合、既存の類似技術の中で最も高速な技術「SASH」の検索時間が、およそ2.4ミリ秒かかるのに対して、「NGT」は最速でおよそ0.6ミリ秒と約4.0倍の高速性を実現しました。なお、既存の類似技術の中でも主流の技術「FLANN」と比べると、約12.3倍（およそ7.4ミリ秒）も高速であることも証明されました。詳細は、以下論文をご参照ください。

Kohei Sugawara, Hayato Kobayashi, and Masajiro Iwasaki, "On Approximately Searching for Similar Word Embeddings", In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL 2016), pages 2265-2275. Association for Computational Linguistics, 2016.
http://research-lab.yahoo.co.jp/nlp/20160828_sugawara.html

【画像特徴データ研究のエビデンス】
検索精度（適合率）90%として、1,000万件の画像特徴データを対象に検索を行った場合、既存の類似技術の中で最も高速な技術「直積量子化手法（PQ：Product Quantization）」の検索時間が、およそ7.9ミリ秒かかるのに対して、「NGT」は最速でおよそ1.4ミリ秒と約5.6倍の高速性を実現しました。なお、既存の類似技術の中でも主流の技術「FLANN」と比べると、約13.5倍（およそ18.9ミリ秒）も高速であることも証明されました。詳細は、以下論文をご参照ください。
Masajiro Iwasaki, "Pruned Bi-directed K-nearest Neighbor Graph for Proximity Search", In Proceedings of Similarity Search and Applications (SISAP 2016), pages 20-33, 2016
http://link.springer.com/chapter/10.1007/978-3-319-46759-7_2



既存の類似技術との比較グラフ（1千万件の画像特徴データを対象に検索を行った場合）

■「NGT」の応用例
高次元データの高速検索が可能になると、大きく二つの応用可能性が考えられます。
一つ目の応用例は、テキストや画像、商品・ユーザーデータなどの高次元データにおいて、近似したデータのマッチングの高速化を通じて、AI（人工知能）技術のさらなる精度向上に貢献することです。Yahoo! JAPANにおいては、まだ実用化にはいたっていませんが、機械学習やディープラーニングを活用している“スマートフォン用のYahoo! JAPANアプリのニュースなどコンテンツのパーソナライズ配信”や“運用型ディスプレイ広告（Yahoo!ディスプレイアドネットワーク）における広告配信”のさらなる精度向上への応用が考えられます。
二つ目の応用例は、多くの項目があり、フォーマットも入力方法もバラバラな大量のデータを高速に照合することを通じて、データの名寄せなど、企業内にたまっているが活用しきれないビッグデータを活用しやすい形に置き換えるといった、AI（人工知能）活用に求められる質の高いビッグデータ生成に貢献することです。
なお、NGTを活用したスマートフォンアプリとして、スマートフォンのカメラを任意の商品に向けてかざすだけで、「Yahoo!ショッピング」内のさまざまなストアで取り扱われている複数の商品ページの中から、最安値で扱うストアの商品ページを特定できる「サイヤスカメラ」

（iOSアプリのみ）を開発しました。11月18日（金）より「Yahoo!ラボ（※）」のスマートフォンアプリとして実験的に公開しています。

iOSアプリ「サイヤスカメラ」ダウンロードページURL：
https://itunes.apple.com/jp/app/saiyasukamera/id1173810477

※Yahoo!ラボ：Yahoo! JAPANの実験的なプロダクト（サービス・機能・仕組み）を、みなさんに体験していただく場として、提供しているページです。
http://labs.yahoo.co.jp/

■OSS公開の狙い
　Yahoo! JAPANは、メディア・コマース・決済などにおいて国内トップクラスのユーザーを抱えるサービスを多数提供しており、その裏側でさまざまな種類のビッグデータが発生しています。
　このような「マルチビッグデータ」を保持している企業は、世界的にも稀有な存在であり、近年、ますますデータの重要性が高まる中、国内外のデータサイエンティストや企業より、注目を集めています。
　また、「マルチビッグデータ」を保持するだけでなく、その利活用のために「技術で世界TOP10」を掲げ、データサイエンス領域における研究開発を推し進めています。
　研究開発の特徴としては、大学や研究機関との共同研究やOSSコミュニティへの貢献強化など、「オープン」なコラボレーションを推進しており、今回のOSS公開もその流れをふまえた、さらなる先端研究の発展のために行う取り組みの一つです。
公開後もOSSコミュニティのコミッターとして、コミュニティ活性化に向けてノウハウの提供などを行う予定です。

　Yahoo! JAPANでは今後も、「技術で世界TOP10」を掲げ、世界的にも稀有な「マルチビッグデータ」の利活用を進めていくことで、人々や社会のさまざまな「課題」を解決してまいります。

■GitHub 内のNGT公開ページのアドレス　https://github.com/yahoojapan/NGT