

検索意図とタスクマイニング: NTCIR-11 IMine タスクへの誘い

Intent and Search Task Mining: Invitation to the NTCIR-11 IMine Task

山本岳洋^{1*} Yiqun Liu² Ruihua Song³ Min Zhang²
Zhicheng Dou³ 加藤誠¹ 大島裕明¹ Ke Zhou⁴

¹ 京都大学

¹ Kyoto University

² 清華大学

² Tsinghua University

³ マイクロソフトリサーチアジア

³ Microsoft Research Asia

⁴ エディンバラ大学

⁴ University of Edinburgh

Abstract: This paper introduces the NTCIR-11 IMine task, whose aim is to explore and evaluate technologies of mining and satisfying different user intents behind a Web search query. The NTCIR-11 IMine task comprises three subtasks: (1) Subtopic Mining, (2) Document Ranking, and (3) Search Task Mining. We briefly introduce these subtasks and how to participate in them. Japanese researchers in information retrieval, natural language processing, interactive information access, or machine learning are highly welcome to participate in the IMine task to accelerate their research.

1 はじめに

NTCIR¹ (NII Testbeds and Community for Information access Research) は国立情報学研究所が主催する、情報アクセスに関する評価型ワークショップである。NTCIR は 1999 年にスタートし、米国の TREC² (Text REtrieval Conference) や欧州の CREF³ (Cross-Language Evaluation Forum) などと同様に、複数の参加者が、同じ目的、テストコレクションのもと、互いに技術を競い合う、もしくは協力することを通じて情報アクセス技術を進歩させることを目的とした国際的なワークショップである。2013 年 6 月に開催された NTCIR-10 では、7 つのタスクが開催され、13 の国と地域から計 103 チームが参加した [3]。NTCIR は 1 年半の周期で開催されており、次回の NTCIR-11 は 2014 年 12 月 9 日から 12 日かけて開催される。

我々は NTCIR-11 にむけて、IMine (Intent and Search

Task Mining) というタスクをオーガナイズしている。IMine タスクは、NTCIR-9 および NTCIR-10 でそれぞれオーガナイズされた、INTENT [6] と INTENT-2 [4, 5] タスクがその前身となっており、INTENT-2 では、日本、中国、韓国、フランスの 4 カ国から計 11 チームの参加があった。INTENT、INTENT-2、そして IMine タスクの目的は、ウェブ検索ユーザの入力したクエリから、その背後にある多様な検索意図を発見し、異なる検索意図を持ったユーザを満足させる方法、およびその適切な評価手法を確立することである。

本稿では、2014 年 12 月にかけて開催される IMine タスクの概要とその参加方法について述べる。これまでの INTENT および INTENT-2 では、日本からの参加チームが多いとはいえなかった。そこで、日本における、情報検索、対話的情報アクセス、自然言語処理、機械学習など、さまざまな分野の研究者の方々に、本稿をご参考にしていただき、是非 IMine タスクに参加していただきたい。

*連絡先：京都大学大学院情報学研究科
〒 606-8501 京都府京都市左京区吉田本町
E-mail: tyamamot@dl.kuis.kyoto-u.ac.jp

¹<http://research.nii.ac.jp/ntcir/index-ja.html>

²<http://trec.nist.gov/>

³<http://www.clef-initiative.eu/>

2 IMine: 検索意図とタスクマイニングに関するタスク

2.1 タスクの概要

ウェブ検索エンジンを利用するユーザの検索意図は多様であるにもかかわらず、検索エンジンに投入されるクエリは短かったり、曖昧であったりすることが多い [7]。そのため、クエリがユーザの情報要求を必要十分な形で表現できていることは稀である。たとえば「ハリーポッター」というクエリは、ハリーポッターの映画に関する情報を求めている場合もあれば、ハリーポッターの書籍、あるいは主人公に関する情報を求めている場合もあり、適合する文書はユーザの検索意図によって異なる。

このような、異なる検索意図を持ったユーザを満足させるための技術として、情報検索の分野では近年、検索結果多様化 (search result diversification) に関する研究が盛んである [1]。検索結果多様化に関する研究で重要となる技術は、与えられたクエリ (トピック) に対して、どのような検索意図が存在するのかを発見し、それにもとづき、多様性を考慮しながら文書をランキングすることで、上位の検索結果だけでできるだけ多くのユーザを満足させることである。

今回行う IMine タスクでは、以下の 3 つのサブタスクを設けている。

- サブトピックマイニング
- ドキュメントランキング
- 検索タスクマイニング

このうち、サブトピックマイニングとドキュメントランキングサブタスクに関しては、INTENT および INTENT-2 でも開催していたサブタスクであり、検索タスクマイニングサブタスクは今回 IMine で初めて取り組むサブタスクである。IMine タスクでは、表 1 に示すように日・英・中の 3 つの言語でのタスクを用意しており、参加者はどの言語のサブタスクでも参加できる。

2.2 サブトピックマイニング

サブトピックマイニング (Subtopic Mining) サブタスクは、与えられたクエリ (トピック) に対し、考えられる意図を文字列で表現したもの (サブトピック) を重要度でランクづけして出力するタスクである。たとえば「ハリーポッター」というクエリに対し、「ハリーポッター 本」、「ハリーポッター 映画」、「ハリーポッ

表 1: IMine で実施されるサブタスク

サブタスク	実施される言語		
	日本語	英語	中国語
サブトピックマイニング	x	x	x
ドキュメントランキング		x	x
検索タスクマイニング	x		

ター 主人公」といったサブトピックが出力として考えられる。

どのような情報源を用い、どのようなアプローチでサブトピックを抽出するかは参加者に委ねられている。たとえば、これまで行ってきた INTENT および INTENT-2 タスクでは、複数の検索エンジンのクエリ推薦や Wikipedia の曖昧性解消ページを用いて単語をクラスタリングすることで、重要なサブトピックを発見する手法が良い成績を収めている。また、我々オーガナイザも、参加者に対して、商用検索エンジンのクリックスルーデータ (中国語) やクエリ推薦 (日本語, 中国語, 英語) を情報源として提供することを予定している。

今回 IMine タスクで新しく取り組む試みとして、これまで、システムはサブトピックを一次元のリストとして出力することが求められていたのに対して、IMine タスクでは、2 次元の階層的なリストを出力することを予定している。サブトピックを階層的に抽出することで、多様な検索意図をより詳細に理解できると考えている。

2.3 ドキュメントランキング

ドキュメントランキング (Document Ranking) サブタスクは、TREC のタスクと同様、ウェブ検索結果を多様化するタスクである。このサブタスクでは、サブトピックマイニングと同一のクエリに対して、文書を適合性と多様性に基づいてランクづけして出力することが求められる。この際、ランキング対象となる文書集合として、参加者に対してあらかじめクロールされたウェブページのデータセットが提供される。そのため、参加者自らがウェブページをクロールする必要はなく、与えられた文書集合をランキングするだけで良い。

これまで行ってきた INTENT および INTENT-2 タスクでは、クエリ推薦やウェブページのドメイン、アンカーテキストなどの異種情報源を統合して検索結果を多様化する手法 [2] やクエリの意図を考慮し、選択的に検索結果の多様化を行う手法 [8] などが好成績を収めている。

サブトピックランキングおよびドキュメントランキングサブタスクの、各参加チームの具体的なアプローチについては、NTCIR-10 のオンライン論文集⁴より閲覧できる。

2.4 検索タスクマイニング

検索タスクマイニング (Search Task Mining) サブタスクは、IMine で今回新しく取り組むサブタスクである。本サブタスクは、2.2 節で述べたサブトピックマイニングと類似のタスクである。サブトピックマイニングが、与えられたトピックのサブトピックを抽出するのに対して、検索タスクマイニングでは、与えられたクエリに対し、クエリの背後にある目的を達成するために必要な行動 (タスク) を、文字列で表現したものを出力するタスクである。たとえば、「花粉症対策」というクエリに対し、「マスクをつける」、「アレロック錠剤を摂取する」、「レーザー治療を受ける」、「花粉症に効くツボを押す」といった出力が考えられる。

検索タスクマイニングでは、サブトピックマイニングが重要度でランクづけされたリストのみを考慮するのに対して、検索タスクマイニングでは、重要度だけでなく、得られたタスク間の順序関係についても出力する必要がある。たとえば、「引越手続きをする」というクエリに対しては、「転入届けを提出する」、「転出届けを提出する」などのタスクが出力として考えられるが、このとき「転入届けを提出する」というタスクは「転出届けを提出する」というタスクを実行する前に実行する必要がある。

このように、サブトピックマイニングが検索意図の「トピック」に注目していたのに対して、検索タスクマイニングでは、検索ユーザの「行動」に着目したタスクである。なお、本サブタスクは京都大学の3名によりオーガナイズされ、日本語で実施される予定である。

2.5 IMine@NTCIR-11 にご参加ください

IMine タスクに参加するためには、まず、2013 年 12 月 20 日までに、NII の NTCIR-11 のホームページ上から参加登録をする必要がある。その後は、表 2 に示した予定で、参加者はシステムの作成、結果の出力、論文執筆、会議への参加などを進めることとなる。また、検索タスクマイニングサブタスクに関しては、一部スケジュールが異なっており、表 3 に従って進められる予定である。IMine および検索タスクマイニングサブ

表 2: IMine のスケジュール (2013 年 10 月時のものです)。

2013 年 12 月 20 日	NTCIR-11 への参加登録
2014 年 1 月	評価用クエリを参加者に配布
2014 年 3 月	システム出力結果提出
2014 年 8 月	評価結果を参加者に配布
2014 年 9 月	論文初稿締め切り
2014 年 10 月	論文最終稿締め切り
2014 年 12 月 9 日 ~ 12 日	NTCIR-11@NII

表 3: 検索タスクマイニングサブタスクのスケジュール (2013 年 10 月時のものです)。

2013 年 12 月 20 日	NTCIR-11 への参加登録
2014 年 2 月	サンプルクエリを参加者に配布
2014 年 3 月	評価用クエリを参加者に配布
2014 年 4 月	システム出力結果提出
2014 年 8 月	評価結果を参加者に配布
2014 年 9 月	論文初稿締め切り
2014 年 10 月	論文最終稿締め切り
2014 年 12 月 9 日 ~ 12 日	NTCIR-11@NII

タスクに関する、より詳細な情報は、それぞれのウェブページ⁵⁶にアクセスいただきたい。

IMine タスクでは、日本語で実施されるサブタスクはサブトピックマイニングと検索タスクマイニングの2種類であるが、両サブタスクとも大規模なクロールやインデックスを必要としないため、手軽に参加していただけたと考えている。情報検索、ウェブマイニング、対話的情報アクセス、自然言語処理や機械学習といった、さまざまな分野の研究者に是非参加していただき、情報アクセスに関する研究を、ともに協力して進めていきたい。

3 むすび

本稿では、NTCIR-11 で行われる IMine タスクについて紹介した。さまざまな分野の研究者、特に日本の研究者の参加には是非本タスクに参加していただき、研究に役立てていただきたい。

謝辞

INTENT, INTENT-2 の各オーガナイザおよび参加チーム、NTCIR の各チェア・事務局の日頃のご尽力に感謝いたします。

⁴http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings10/NTCIR/toc_ntcir.html

⁵IMine タスク, <http://www.thuir.org/IMine/>

⁶検索タスクマイニングサブタスク, <http://www.dl.kuis.kyoto-u.ac.jp/ntcir-11/taskmine/>

参考文献

- [1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying search results. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, pages 5–14, 2009.
- [2] Z. Dou, S. Hu, K. Chen, R. Song, and J.-R. Wen. Multi-dimensional search result diversification. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 475–484, 2011.
- [3] H. Joho and T. Sakai. Overview of NTCIR-10. In *Proceedings of NTCIR-10*, 2013.
- [4] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, M. P. Kato, R. Song, and M. Iwata. Summary of the NTCIR-10 INTENT-2 task: Subtopic mining and search result diversification. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 761–764, 2013.
- [5] T. Sakai, Z. Dou, T. Yamamoto, Y. Liu, M. Zhang, and R. Song. Overview of the NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.
- [6] R. Song, M. Zhang, T. Sakai, M. P. Kato, Y. Liu, M. Sugimoto, Q. Wang, and N. Orii. Overview of the NTCIR-9 INTENT task. In *Proceedings of NTCIR-9*, 2011.
- [7] J. Teevan, S. Dumais, and E. Horvitz. Potential for personalization. *ACM Transactions on Computer-Human Interaction*, 17(1):4:1–4:31, 2010.
- [8] Y. Xue, F. Chen, A. Damien, C. Luo, X. Li, S. Huo, M. Zhang, Y. Liu, and S. Ma. THUIR at NTCIR-10 INTENT-2 task. In *Proceedings of NTCIR-10*, 2013.