

コンピュータはどのくらいうまく情報を探せるのか

NTCIR(えんていさいる)-情報アクセス技術の比較評価基盤と研究コミュニティ

How Computers Can Effectively Search Information

Leveraging Information Access Technologies by Community-Based Evaluation in NTCIR

神門 典子 (Noriko Kando) 相澤 彰子 (Akiko Aizawa) 宮尾 祐介 (Yusuke Miyao) Teruko Mitamura (CMU/NII)、武田浩一 (Koichi Takeda, IBM/NII), Douglas Oard (UMD/NII), Mark Sanderson (RMIT/NII), 上保 秀夫 (Hideo Joho, 筑波大/NII) 世界中から48人の研究者

何がわかる？

- 斬新な研究をするための研究基盤(土台)を提供します
- 研究がスピードアップします
- 評価実験が容易になり、検証→改善のサイクルが早まります。多チームの多様なアプローチから相互に学び合えます
- グランドチャレンジのための基盤とマイルストーンを提供します
- よりよい検索システムを実現するために、研究コミュニティの形成を支援しています。

どんな研究？

17カ国の120以上の研究チームと一緒に、情報検索、検索意図マイニング、要約、質問応答、含意の推定、情報抽出、リンク発見、テキストマイニング、意見分析、機械翻訳などの有効性を検証するために、実験データセットの構築、比較評価、テスト手法の研究しています。

研究状況

NTCIR NII Testbeds and Community for Information access Research

情報検索・アクセス技術の評価と性能比較のための研究基盤 NTCIR (エンティサイル)

目的: 大規模な評価実験用の研究基盤を提供することによって情報アクセス技術研究の促進

- 基盤: データ、評価手法、コミュニティ
- システム間の性能比較、手法の特性の相互比較 技術転移 互いに学びあう場

1997年末にプロジェクト開始

- 一年半毎に成果報告会を開催

再利用可能な大規模データセット(テストコレクション)

- 学術、新聞記事、特許、Web、ヤフー知恵袋、Wikipedia、大学入試など
- 研究目的利用で公開

研究部門(タスク)

- 情報検索: 言語横断、特許、Web
- 質問応答: 単言語、言語横断
- 要約、動向情報、特許マップ自動生成

これまでの参加登録チームの数: 合計 963 チーム

テストコレクション 利用申請者: 合計 3543名 現在利用中: 2347名

情報検索の歴史(情報処理学会50年史より)

情報検索の歴史(情報処理学会50年史より)

1950年代: クラウド型実験

1960年代: ベクトル空間型 SMART

1970年代: プーリング法 IBM/STARS

1980年代: 確率論 Blai-Saltan

1990年代: 大規模TCで検索可能に 利用者のタスクに依存 多様な利用者のタスクに依存 多様な利用者のタスクに依存

2000年代: 大規模TCで検索可能に 利用者のタスクに依存 多様な利用者のタスクに依存

2010年代: 大規模TCで検索可能に 利用者のタスクに依存 多様な利用者のタスクに依存

NTCIRのこれまでの歩み

NTCIR成果報告会の様子

NTCIRワークショップの準備

- ・ 研究部門(タスク)のプロポーザル募集
- ・ 研究部門・対象技術の選定 - プログラム委員会 - データ提供機関との協議・契約
- ・ 研究部門ごとに、実験・評価手法の提案・議論
- ・ 参加申込
 - 訓練用データ(文書、課題、正解)配布
 - 各参加者が実験、チューニング
 - 評価用データ(文書、課題)配布
- ・ 検索結果の提出
- ・ 提出結果から正解検索率、人手判定
- ・ 正解公開、評価結果発表
- ・ 成果報告会 次回についての協議

グランドチャレンジ「ロボットは東大に入れるか」と連携

2011年 再融合へ「ロボットは東大に入れるか？」

2011年 キックオフ: 社会系センター試験に6チーム16システムがチャレンジ!

2012年の目標: 社会系科目: 教科書・Wikipediaを検索し、推論して答える

今後は...

NTCIRの活動概要

- テストコレクション(実験用データセット)50種類を公開
- NTCIR成果報告会(国際ワークショップ)9回 (10日目前準備中)
- 国際ジャーナル特集号6回、国際学会のワークショップ8回
- 国際会議における招待講演 50回以上
- Google ScholarでNTCIRに言及している論文: 4210件
 - 高Citation count論文も注目!
- 米国政府大規模研究助成プログラムの指定評価 AQUAINT など

今までの主な参加団体:

カリフォルニア大バークレイ、カーネギーメロン大、コーネル大、メリランド大、スタンフォード大、IBM、MIT、BBN、SIGS、XEROX、Dublin City U、IBM、Microsoft、ヤフー、RMIT、ケンブリッジ大、グラスゴー大、RWTH、韓国KAIST、国立シンガポール大、国立台湾大、I2R、中国科学院、清華大、NICT、産総研、東大、京大、北大、奈良先端大、北陸先端大、筑波大、豊橋技科大、長岡技科大、NTT、日立、東芝、松下、リコー、FujiXEROX、JustSYSTEM、富士通研、NTTデータ、NEC、沖電気、など

NTCIR-10

参加チーム募集中!

| NTCIR 回次 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 研究部門(タスク) |
|-----------|-----|-----|-----|-----|-----|-----|-----|-----|-----|-----|----------------|
| 技術分野 | '99 | '01 | '02 | '04 | '05 | '07 | '08 | '10 | '11 | '13 | |
| ソーシャルメディア | | | | | | | | | | | コミュニティQA |
| IGCC | | | | | | | | | | | 意見分析 |
| モジュール評価 | | | | | | | | | | | 検索+推論 |
| | | | | | | | | | | | 質問応答+検索 |
| | | | | | | | | | | | 入試問題 |
| | | | | | | | | | | | 音声文書検索 |
| | | | | | | | | | | | 地理・時間情報検索 |
| | | | | | | | | | | | 特許検索 |
| | | | | | | | | | | | あらゆるタイプの質問 |
| 質問応答 | | | | | | | | | | | 対話 |
| | | | | | | | | | | | 言語横断 |
| | | | | | | | | | | | 事実を導く、リスト型 |
| 情報抽出と検索 | | | | | | | | | | | リンク発見 |
| 意味処理 | | | | | | | | | | | 推論 |
| 要約・統合 | | | | | | | | | | | テキストマイニング・分類 |
| | | | | | | | | | | | 動向情報の抽出・可視化 |
| | | | | | | | | | | | テキスト要約 |
| インタラクティブ | | | | | | | | | | | インタラクティブ検索と可視化 |
| Web | | | | | | | | | | | 検索意図・多様な意図の検索 |
| | | | | | | | | | | | 検索結果一見ページの質 |
| | | | | | | | | | | | Web検索 |
| | | | | | | | | | | | 機械翻訳 |
| 言語横断 | | | | | | | | | | | 言語横断検索 |
| | | | | | | | | | | | 英語以外の検索 |
| テキスト検索 | | | | | | | | | | | 日本語検索 |

年は成果報告会開催年。研究部門は18ヶ月前に開始

さまざまなタスク

2011年 キックオフ: 社会系センター試験に6チーム16システムがチャレンジ!

2012年の目標: 社会系科目: 教科書・Wikipediaを検索し、推論して答える

今後は...

NTCIR-10: ねらい

- ・ 強固な運営基盤
 - 運営組織の強化
 - タスクの多様性
 - 情報アクセスにおける幅広いコンテキストを対象
 - 多様なメディアを研究対象に
- ・ コミュニティ主導のタスク運営
 - 研究の持続性
 - 研究資源利用の促進
 - 成果報告会で特別ブースの設置

NTCIR-10: 運営組織

General Co-Chairs: 加藤 博樹 (東京大学), 神門 典子 (NII), Douglas W. Oard (University of Maryland), Mark Sanderson (RMIT), Program Co-Chairs: 藤井 哲也 (MIRA), 上保 秀夫 (筑波大学)

Task Organizers: タスクオーナーは世界中の個人研究者、参加者(あななです!)、EVA 2013 Co-Chairs: Ruitan Song (MSRA), William Webber (University of Maryland)

NTCIR-10の研究部門

情報アクセスのコンテキスト

幅広いコンテキストと多様なメディア

Intent 検索意図

Click ユーザのクリック

RITE 含意の推定

Math 数式検索

PatentMT 特許翻訳

SpokenDoc 音声検索

CrossLink リンク発見

重要な日程

2013.03.08 キックオフイベント

2013.06.30 タスク参加登録締切

2013.07.01 実験データ配布開始

2013.07.20-2013.12 予備テスト

2013.09.20-2013.02 ノート

2013.02.01 評価結果発送

2013.02.01 タスク結果発表一部公開

2013.02.01 成果報告会論文提出締切

2013.02.01 成果報告会論文発表開始締切

2013.06.18-21 第10回NTCIR成果報告会(東京)

現在、参加団体募集中!

まとめ

- ・ NTCIR-10の目的のサイクル
- ・ 運営組織も進化!
- ・ 7つの魅力的なタスク
- ・ 世界中の個人研究者によって運営
- ・ 斬新な研究をする土台を準備
- ・ 多様なコンテキストとメディア

皆さんの参加をお待ちしています!

参加のメリット (学生編)

- ・ すぐに始められる
- ・ 実績
- ・ 一他参加者との比較による公正な議論
- ・ 成功のステップについて国際コミュニティと意見交換
- ・ 多様なタスク
- ・ 新しい課題への挑戦もあり、それまでの応用もあり

参加のメリット (企業編)

- ・ 開発の促進
- ・ 自社製品のブラッシュアップやバグ取りが短期間で集中的にできる
- ・ 研究資源への早期アクセス
- ・ タスクで構築された様々なリソースを一般公開より先に利用可能

2012年の目標: 社会系科目: 教科書・Wikipediaを検索し、推論して答える

今後は...

- ・ 評価への個別的なニーズへの対応
- ・ 探索的・対話的な情報アクセス
 - 利用者の意図
 - 経緯・知識・リテラシー
- ・ SWIRL 2012で議論された重要トピック
 - 対話検索
 - モバイル検索
 - ゼロクエリ検索

NTC-10 運営メンバー

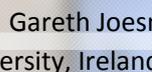
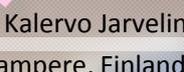
CrossLink



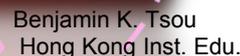
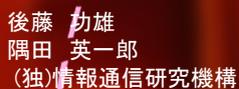
INTENT

1CLICK

General Co-chairs



PatentMT



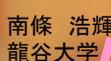
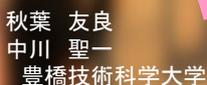
Program Co-chairs



EVIA Co-Chairs



SpokenDoc-2



RITE-2



MATH

