Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

曖昧なクエリと (不) 明快なクエリ: NTCIR-10 INTENT-2 と 1CLICK-2 タスクへの誘い

酒 井 哲 也[†1]

NTCIR (エンティサイル) は国立情報学研究所 (NII) が主催する情報アクセス技術の評価型ワークショップシリーズで，1999 年以来一年半サイクルで開催されている．第 9 回ワークショップ (NTCIR-9) が 2011 年 12 月に終わり，第 10 回 (NTCIR-10) の採択タスクが今年 1 月にアナウンスされたところである．本稿ではそのうち特に INTENT-2 および 1CLICK-2 というタスクについて紹介する．各タスクについて，まず NTCIR-9 での動向について触れ，次に NTCIR-10 に向けた最新情報を提供する．情報検索および自然言語処理の研究者の方々にはこれらのタスクに是非参加していただき，研究の加速に役立てていただきたい．

Ambiguous, Underspecified and Clear Queries: Invitation to NTCIR-10 INTENT-2 and 1CLICK-2 Tasks

Tetsuya Sakai[†1]

NTCIR is a sesquiannual evaluation workshop series on information access which started in 1999 and is run by National Institute of Informatics (NII). The ninth workshop (NTCIR-9) was concluded in December 2011 and the accepted tasks for the tenth workshop (NTCIR-10) have been announced in January 2012. This paper introduces two accepted tasks called INTENT-2 and 1CLICK-2: I first briefly summarise what happened at NTCIR-9 for the initial rounds of these tasks (INTENT-1 and 1CLICK-1), and then provide the latest information on INTENT-2 and 1CLICK-2. Researchers in information retrieval and natural language processing are welcome to participate in these tasks to accelerate their research.

†1 Microsoft Research Asia. tetsuyasakai@acm.org

1. はじめに

NTCIR (エンティサイル) はもともと NII Test Collection for Information Retrieval systems として 1999 年にスタートしたが，最近では NII Testbeds and Community for Information access Research とも呼ばれており，狭義の情報検索 (文書検索) のためのテストコレクション作成を伴う評価型ワークショップから，広義の情報検索 (ユーザを所望の情報に迅速かつ的確に導くための多様な技術，すなわち情報アクセス) の評価型ワークショップへと発展してきた[18]．米国の TREC (Text Retrieval Conference; 1992 年スタート) や欧州の CLEF (Cross-Language Evaluation Forum; 2000 年スタート) などと共に，一定期間内に同じ土俵の上で技術を競い合う，もしくは協力することを通じて情報アクセス技術を振興するための国際的な場となっている．2011 年 12 月に開催された第 9 回 NTCIR ワークショップ (NTCIR-9) には，7 つのタスク (オリンピックで言えば種目に相当する) が設けられ，14 の国もしくは地域から 90 の異なる研究機関が参加した．NTCIR は一年半サイクルで動いているので，次回の NTCIR-10 は 2013 年 6 月に開催される．これに向けて，NTCIR-10 の採択タスクが 2012 年 1 月に公表され，各タスクが始動したところである．

筆者は，NTCIR 全体のプログラム共同議長を務めるとともに，NTCIR-10 に向け採択された 7 つのタスクのうちの INTENT-2[⋆1] と 1CLICK-2[⋆2]という 2 つのタスクのオーガナイザでもある[⋆3]．前者はウェブ検索における曖昧もしくは不明快なクエリを扱うタスク，後者はモバイル検索などの状況でユーザの情報要求を瞬時に満たすことを狙ったタスクである．しかし，前回の NTCIR-9 における日本からの参加チームは第 1 回 INTENT タスク (INTENT-1) で全 17 チーム中 1 チーム，第 1 回 1CLICK (INTENT-1 の試験的サブタスクとして開催) で全 3 チーム中 2 チームと，大変少なかった[⋆4]．両タスクは NTCIR-10 採択タスクのうち狭義の情報検索と最も密接なタスクなので，特に日本における情報検索・言語処理研究者にもっと多く参加してもらい，研究の加速に役立てていただきたい．そこで本稿では，NTCIR-9 における INTENT-1 および 1CLICK-1 の概要を報告し，これから 2013 年 6 月にかけて実行される INTENT-2 および 1CLICK-2 の最新情報を提供する．

なお，今回の研究発表会では，「NTCIR-9 総括と今後の展望」という発表が別途あるの

⋆1 http://research.microsoft.com/en-us/people/tesakai/intent2.aspx
⋆2 http://research.microsoft.com/en-us/people/tesakai/1click2.aspx
⋆3 筆者は NTCIR-10 プログラム委員会によるタスク選定の際，投票を棄権した．
⋆4 NTCIR-9 全体としては，全 90 の異なる参加チームのうち，日本は 34 チームで最多であった．

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

で[19]，INTENT と 1CLICK 以外のタスクについてはそちらを参照いただきたい．

2. ユーザ意図に着目した検索タスク INTENT

2.1 INTENT-1@NTCIR-9 概観

ウェブ検索エンジンに入力されるクエリが，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．「オフィス」というクエリは，マイクロソフトの製品を意図しているのか，仕事場を意図しているのか (**曖昧なクエリ**)．また，「ハリーポッター」というクエリは，本を意図しているのか，映画を意図しているのか，主人公のキャラクターを意図しているのか (**不明快なクエリ**)．

上記のようなタイプのクエリに対する技術として，最近 search result diversification (**検索結果多様化**) の研究が盛んである．従来の情報検索が relevance (**適合性**) のみに基づきシステムを最適化しようとしていたのに対し，多様化の研究では、relevance と diversity (**多様性**) の双方を最大化しバランスをとることが要求される．例えば，「ハリーポッター」というクエリに対し，前述の様々な意図に適合するウェブページをうまく織り交ぜて提示し，上位の検索結果だけでなるべく多くのユーザ意図を満たすことを目指す．前述の TREC では，英語を対象としたウェブ検索結果多様化タスクが 2009 年より毎年開催されている．

NTCIR-9 でスタートした INTENT タスクは，上述の TREC のタスクと似ているが，日本語および中国語を対象としており，TREC にはないいくつかの試みを行っている．例えば，INTENT は **Subtopic Mining** と **Document Ranking** という 2 つのサブタスクにより構成されているが，前者は TREC にはないものである[⋆1]．

2.1.1 INTENT-1 のタスク仕様

Subtopic Mining サブタスクは，与えられたクエリ (例:「ハリーポッター」) 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ものである (例:「ハリーポッター 本」「ハリーポッター 映画」「ハリーポッター 主人公」...)．サブトピックをどのような知識源から抽出するかは参加者の自由である．例えば，ウェブ検索エンジンをもつ研究機関であればセッションやクリックスルーのデータ，そうでなければウェブ検索エンジンが出力する query suggestion などの利用が考えられる[⋆2]．

一方，Document Ranking サブタスクは，TREC のタスクと同様ウェブ検索結果の多様化を狙ったものである．参加システムへの入力は Subtopic Mining と同一のクエリであるが，要求される出力はウェブ検索結果，すなわちランクづけされたウェブページである．なお，検索対象となるのは予めクロールされた中国語 1.38 億ページもしくは日本語 0.67 億ページのウェブデータで，評価の対象となるのは検索結果の上位 10-30 件である．前述のとおり，検索結果多様化は上位の検索結果により多くのユーザ意図を満足させることを狙ったものである．

図 1 (左) に INTENT タスクの流れを示す．タスクオーガナイザは，Subtopic Mining 参加者に，日中それぞれ 100 件の**トピック** (クエリ)[⋆3] を配布する (矢印 1)．参加者は，各トピックに対してサブトピックのランクつきリストを生成し，これらを指定の形式に従いひとつのファイルにまとめたもの (**ラン**) を締切までにオーガナイザに提出する (矢印 2)．オーガナイザは，各トピック毎に，全参加者から提出されたサブトピックの候補を統合し (このデータを**プール**と呼ぶ)，これを評価作業者に渡す (矢印 3)．評価作業者は，同一の内容を意味すると思われるサブトピックを人手で intent (**意図**) にまとめあげる．これにより，各トピックに 1 つ以上の intent が付与される (矢印 4)．評価作業者はさらに，トピックに対する各 intent の重要度を投票する (矢印 5,6)．以上の過程により、各トピックにひとつ以上の intent が，多数決に基づき算出された確率つきで付与される．

一方，タスクオーガナイザは，Document Ranking 参加者についても同一のトピック集合を配布する (矢印 7)．参加者は，各トピックに対してウェブページ ID のランクつきリストを作成し，ランを締切までにオーガナイザに提出する (矢印 8)．オーガナイザは，トピック毎にランを統合したプールを作成し，評価作業者に渡す (矢印 9)．評価作業者は，プール内の各ウェブページを閲覧し，高適合，適合，不適合のラベルを intent 毎に付与する (矢印 10)．すなわち，従来の情報検索における適合性判定のようにページがクエリ「ハリーポッター」に対し適合するかどうかを判定するのではなく，「本としてのハリーポッター」「映画としてのハリーポッター」それぞれの意図に適合するかどうかを判定する．最後に，評価作業者 2 名の判定結果を統合することにより，$L0$ (不適合) から $L4$ (最も高いレベルの適合) の 5 値の適合性判定データが得られる．

なお，TREC の検索結果多様化タスクでは，各 intent は等確率であると仮定されており，

⋆1 TREC 2012 では，英語の Subtopic Mining が導入されるそうである．

⋆2 例えば大学などの研究機関が，商用検索エンジンの query suggestion などをブラックボックス的に使用することの意義については議論の余地があろう．特に，query suggestion は随時変化する可能性があるので，実験の再現性に問題が生じる．この点については 2.2 における INTENT-2 の取り組みを参照いただきたい．

⋆3 トピックは評価タスクにおける検索課題，クエリは検索エンジンへの入力文字列を指す．INTENT タスクでは両者は同一である．

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

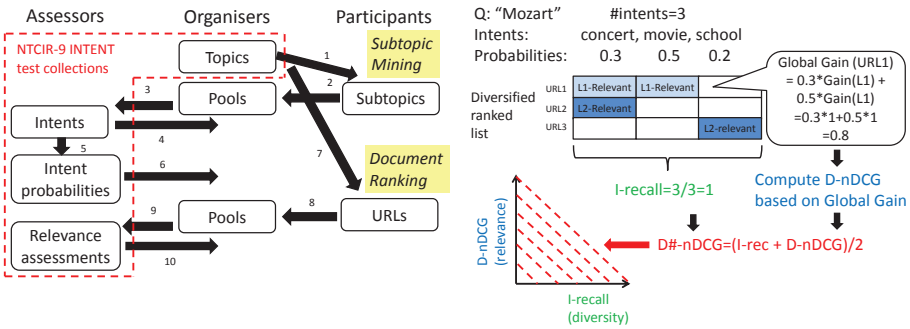

図 1 NTCIR-9 INTENT タスクの流れ (左) / 評価指標 D♯-nDCG の概念図 (右).

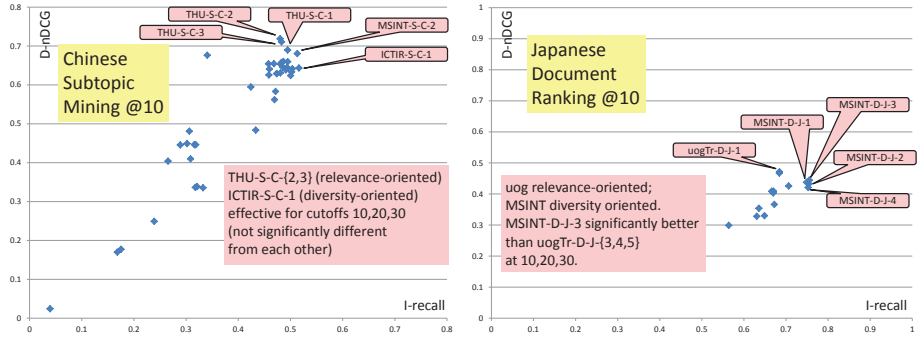

図 2 NTCIR-9 INTENT 中国語 Subtopic Mining サブタスクの公式結果 (左) / 日本語 Document Ranking サブタスクの公式結果 (右).

intent 毎の適合性判定は多値ではなく二値である[3].

2.1.2 INTENT-1 の評価方法

第 1 回 INTENT タスク (INTENT-1) では，Subtopic Mining および Document Ranking において **D♯-nDCG**[22] という同一の評価指標を用い，参加チームの評価を行った．

図 1 (右) に D♯-nDCG による評価の概念図を示す．この例では，Document Ranking サブタスクのトピック「モーツァルト」に対する検索結果を評価している．(実際の言語は日本語もしくは中国語である．) ここで，仮にこのトピックに対しては「モーツァルトの曲の演奏会」「モーツァルトに関する映画」など 3 つの intent が，それぞれ確率 0.3，0.5，0.2 とともに与えられているとする．図中で検索結果 1 位にランクされているウェブページは，最初の 2 つの intent に対してのみ適合しており，かつその適合レベルはともに $L1$ (最も低いレベルの適合) である．例えば適合レベル $L1$ に対し 1 点，$L2$ に対し 2 点，という要領で gain (**利得**，システムに与える褒美[15),16)]) の値を定義する場合，この文書の **global gain** (全 intent を見渡した時の利得) は，図中の吹き出しのように，intent の確率と intent 毎の利得の積の和として簡単に計算できる．

上記 global gain に基づき，**理想的な検索結果**を定義することは容易である．これには，少なくとも 1 つの intent に対し適合と判定された全文書に対して global gain を計算し，これらを global gain の値の降順に並べればよい．このように global gain を用いることにより，従来の情報検索における評価指標，例えば **nDCG**[15),16)] が再定義できる．Global gain に基づく nDCG を **D-nDCG** という[22]．これは，複数の intent とその確率を考慮し，検索結果の適合性を総合的に評価する指標である．

一方，図 1 の検索結果は 3 つの intent を全てカバーしている．このことは，図中にあるように，intent recall (**I-recall**) により評価できる．こちらは，純粋に多様性のみを考慮した評価尺度である．D-nDCG 単体では，多様性が比較的低くても適合性が高いシステムを高く評価する可能性があるため，INTENT タスクでは，D-nDCG と I-recall を平均した D♯-nDCG により参加チームをランクづけしている．これは，図中のグラフのように，参加チームの D-nDCG と I-recall をプロットした場合の，点線で示した等高線に対応する．また，このようなグラフを用いると，参加チームのうちどれが適合性を重視しているか，どれが多様性を重視しているかを把握することが容易である．

D♯-nDCG は，TREC で用いられている **α-nDCG** や **Intent-Aware な評価指標**[1),2)] よりも統計的に信頼性が高く，より直観的な尺度であることが示されている[17),22)]．

なお，Subtopic Mining では，例えば「アマデウス」「モーツァルト 映画」のようなサブトピックが評価作業者により「モーツァルトに関する映画」のような intent にまとめあげられるため，図 1 のウェブページの場合とは異なり，各サブトピックが複数の intent に対応することはない．さらに，図 1 とは異なり，サブトピックは intent に属するか否かの二値情報により管理される．従って，各サブトピックの global gain は，それが属する intent の確率に帰着する．例えば intent「モーツァルトに関する映画」の確率が 0.5 であれば，これに属する任意のサブトピックの global gain は 0.5 となる．

2.1.3 INTENT-1 の公式結果

図 2 に，NTCIR-9 INTENT 中国語 Subtopic Mining サブタスクの公式結果 (左) および日本語 Document Ranking サブタスクの公式結果 (右) の **D-nDCG / I-recall グラフ**

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

を示す．紙面の制約上，日本語 Subtopic Mining および中国語 Document Ranking の結果は割愛する．詳細については NTCIR-9 INTENT の Overview 論文を参照されたい．

図 2 左の中国語 Subtopic Mining では，清華大学 (THU)，中国科学院 (ICTIR)，マイクロソフトリサーチアジア (MSINT) らがよい成績をあげており，これらのシステム間に有意差はない．なかでも THU は適合性指向 (D-nDCG が高い)，ICTIR は多様性指向 (I-recall が高い) となっている．

THU は中国語 Subtopic Mining のために，5 つの検索エンジンの query suggestion と，ウィキペディアの曖昧性解消ページおよびトピック文字列を含むエントリを利用しており，図 2 左が示すようによい成績を収めている．また彼らはこれらの上位のランの他にクリックスルーデータを用いたランも提出している[26]．一方，ICTIR は，クエリログ，オンライン百科事典，3 つの検索エンジンの query suggestion から得たクエリをクラスタリングし，クラスタの大きい順に各クラスタのセントロイドをサブトピックとして提出している[27]．

図 2 右の日本語 Document Ranking では，マイクロソフトリサーチアジア (MSINT)，英グラスゴー大学 (uog) らがよい成績をあげており，前者は多様性指向，後者は適合性指向となっている．また，D♯-nDCG によれば，MSINT と uog のシステムの間には統計的に有意な差がある．

MSINT は日本語 Document Ranking において，マイクロソフトリサーチアジアが提案した多次元の多様化手法[5] を応用している．具体的には，2 種類の query suggestion (検索窓内に表示されるものと検索結果横に表示されるもの) と初期検索結果上位のウェブサイトのドメイン名を 3 つの次元に見立てて多様化を行っているが，3 つの次元の異なる組み合わせによるランの間には有意差はない[6]．一方，uog は，グラスゴー大学が開発した xQuAD[23] という検索結果多様化手法を用いて中国語および日本語検索に挑戦している．図 2 右で示されている彼らの上位のランは，検索エンジン Bing の日本語 query suggestion を利用したものである．この他，適合性と多様性のバランスをクエリ毎に切り替える selective diversification (**選択的多様化**) にも挑戦している[24]．

各参加チームのより具体的な取り組みについては，NTCIR-9 のオンライン論文集[⋆1]をご参照いただきたい．

2.2 INTENT-2@NTCIR-10 にご参加ください!

これから 2013 年 6 月にかけて進められる第 2 回 INTENT タスク (INTENT-2) の，INTENT-1 との主な相違点は以下のとおりである．

- Subtopic Mining のみ，対称言語を中日**英**に拡張予定である．具体的には，英語の検索結果多様化に取り組んでいる TREC の diversity task とトピック集合を共有する．
- トピック集合に，おそらく検索結果の多様化を必要としない **very clear queries** を含める．具体的には，答えがひとつ見つかれば済むクエリを用意する予定である．従って，参加システムは，各トピックについて多様化の必要性を判定した上で検索戦略を切り替えることも可能である．
- オーガナイザが，代表的な検索エンジンの query suggestion データを採集し，これを Subtopic Mining および Document Ranking 参加者に配布する．これにより，実験の再現性と公平性を高める．
- Document Ranking では，多様化されていない検索結果を (ウェブページの中身とともに) baseline として参加者に提供する．これにより，検索対象のウェブデータや検索エンジンをもっていない研究チームでも，baseline を再ランキングするなどの簡単なアプローチによりタスクに参加できる．
- 参加チームには，NTCIR-10 の新しいトピックセットだけでなく，NTCIR-9 のトピックセットに対しても同じシステムによる結果を作成してもらう．これにより NTCIR-9 と NTCIR-10 のトピックの等価性の議論ができる[⋆2]．さらに，NTCIR-9 から参加しているチームには，NTCIR-9 で使用したシステムによる NTCIR-10 トピックセットの処理結果も作成してもらう．これにより技術の進歩について議論できる[14]．

また，参加システムの評価方法についても，例えば navigational な intent (例: 特定のウェブサイトにアクセスしたい) を考慮したもの[17] など，D♯-nDCG に加え新たな方法を模索する予定である．

INTENT-2 のスケジュールは以下のとおりである．

2012 年 5 月	(オーガナイザ) トピックおよび baseline 検索結果を参加者に配布
2012 年 7 月	(参加者) Subtopic Mining および Document Ranking のラン提出締切
2012 年 8-12 月	(判定作業者) intent 作成，適合性判定
2013 年 1 月	(オーガナイザ) 評価結果を参加者に配布
2013 年 3 月	(参加者) 論文初稿締切
2013 年 5 月	(参加者・オーガナイザ) 論文最終稿締切
2013 年 6 月 18-21 日	NTCIR-10@NII

⋆1 http://research.nii.ac.jp/ntcir/workshop/OnlineProceedings9/NTCIR/toc_ntcir.html

⋆2 もちろん，NTCIR-10 参加システムは，既存の NTCIR-9 トピックセットに対しチューニングされている可能性はある．

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

参加者としては，システムを開発し事前準備を行う 5 月までと，トピックが配布されてから結果を提出するまでの 5〜7 月が勝負どころとなる．ただし，提出するランは全て自動処理に基づくものでなければならないので，事前準備をきちんとしておけば，7 月までそれほど労力を費やす必要はないはずである．自前の検索エンジン等を用いランを提出するチームは，中国語や日本語の検索対象ウェブデータを事前に入手し，検索インデクス作成などの準備を行っておく必要がある．

タスクに参加するには，NII が用意する NTCIR-10 のホームページ[⋆1] からオンライン登録すればよい．なお参加チームには，ランおよび論文提出と 2013 年 6 月の会議におけるポスター発表が義務付けられているのでご留意いただきたい．

3. 検索ボタンを押した直後にユーザを満足させるタスク 1CLICK

3.1 1CLICK-1@NTCIR-9 概観

第 1 回 1CLICK (1CLICK-1) は，NTCIR-9 では INTENT のサブタスクという扱いであったが，もともと別個のタスクとして提案された趣の異なるものである．なお対称言語は日本語のみであった．

図 3 (左) に，通常のウェブ検索におけるユーザ操作の流れと，**One Click Access** (1CLICK) におけるユーザの操作の流れの違いを示す．INTENT の Document Ranking サブタスクも，ウェブページのランクつきリストを出力するという意味で，前者に該当する．この図が示すように，1CLICK では，ウェブページのランクつきリストをユーザに提示するのではなく，一般には複数のページから得られる重要情報を短いテキストに集約してユーザに提示することを狙っている．このように，検索ボタンを押した直後に欲しい情報が簡潔な形で得られるという形態の情報アクセスは，特にモバイル環境において有用であると考えられる．なお，1CLICK の枠組みにおけるクエリの種類としては，ユーザ・システム間の煩雑なインタラクションを必要としない，比較的**明快なクエリ**を想定している[⋆2]．

3.1.1 1CLICK-1 のタスク仕様

1CLICK の参加チームは，与えられたクエリに対し，重要な情報を 500 字もしくは 140 字のテキストにまとめた形で出力する．前者はデスクトップ PC 環境で通常のウェブ検索結果を表示させた場合に，スクロールなしで閲覧できる上位数件のスニペットの分量を想定したもの，後者は携帯電話の画面に簡潔に表示する分量 (もしくは 1 ツイート) を想定したものである．このことから，それぞれの字数制限に従ったランを DESKTOP ランおよび MOBILE ランと呼んでいる．

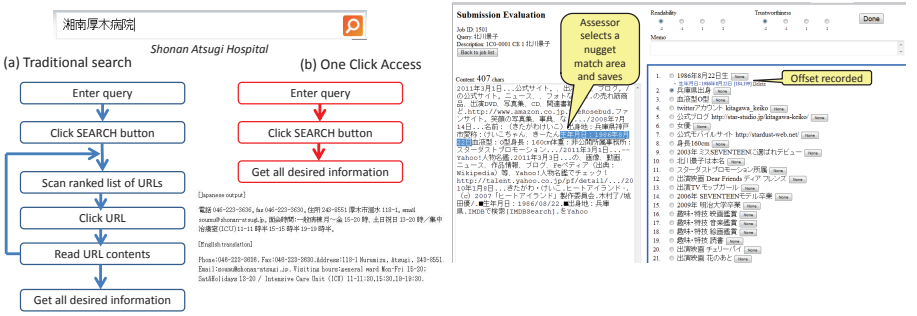

図 3 従来型ウェブ検索とワンクリックアクセスの違い (左) / ナゲット適合評価インタフェース (右).

参加システムには，以下の条件が要求される．

- 重要な情報をなるべく優先して提示．
- ユーザの読むテキスト量を最小化．

1CLICK-1 では，クエリのタイプとして，デスクトップ環境とモバイル環境の英日中クエリログに関する従来研究[8] に基づき CELEBRITY (有名人)，LOCAL (地名)，DEFINITION (定義)，QA (質問応答) の 4 種類を用意した．要求される情報の種類はクエリタイプ毎に設定し，各クエリタイプ 15 件，合計 60 件の評価用クエリを用意した．

参加チームには，検索エンジン API，ウィキペディアのダンプデータ，ヤフー知恵袋のデータなどあらゆるリソースを利用し，重要な情報を収集することを許した．また、提出するランのファイルには，500 字もしくは 140 字の出力テキストに加え，これを作成するのに要したウェブページの URL のリストを付与することを義務づけた．あらゆる知識源を用いてよいという意味で，このようにして生成されたランを **OPEN** ランと呼ぶ．なお，INTENT タスク同様，人手を要するランは許していない．

3.1.2 1CLICK-1 の評価方法

「重要な情報を優先提示し，ユーザの読むテキスト量を最小化する」という要求を満たすシステムを評価するため，1CLICK では**ナゲット** (情報の最小基本単位) に基づく新しい評価方法を設計した[20]．

⋆1 http://research.nii.ac.jp/ntcir/ntcir-10/index.html
⋆2 実際のテストクエリには意図的に曖昧なクエリも混ぜてあるが，曖昧性を考慮した特別な評価方法は今のところ採用していない．

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

ナゲットに基づくテキストの評価自体はテキスト要約[12] および質問応答[4),9)] で従来から行われてきたが，1CLICK における評価のこれらとの違いは，システムの出力に含まれるナゲットの位置情報を考慮する点である．情報検索評価指標 nDCG が文書のランクに基づきその価値を discount (**減損**[15),16)]) させるのと同様に，1CLICK ではナゲットの出現位置 (テキスト先頭からのオフセット値) に基づきその価値を減損させる．

図 3 (右) に，1CLICK 参加システム評価のために開発されたナゲット適合評価インタフェース[21] の使用例を示す．1CLICK では，各クエリに対し，予め正解ナゲット集合を作成しておく．(さらに，各ナゲットには多数決により重要度を付与しておく．) 図中の右側に列挙されているのは，クエリ「北川景子」に対する正解ナゲットの一部である．また，左側に表示されているのは，ある参加システムの出力結果である．評価作業者は，右のナゲットと左のテキストを見比べて，各ナゲットが含まれているか否か，かつ，含まれている場合はどの部分が該当するかをマウスドラッグと左クリックにより指定する．これにより，各該当ナゲットの位置情報が記録される．なお，1CLICK-1 では，上記インタフェースを用いたランの評価作業を参加チームのメンバが分担して行った．

評価尺度としては，位置情報を考慮したナゲットベースの尺度 **S-measure**[20] を用いた[⋆1]．これは，重要度が高く，かつ簡潔に表現されたナゲットを優先して提示するシステムを高く評価する尺度である．S-measure は，ユーザが情報収集に使える時間の上限，もしくはユーザの読むテキスト量の上限を表すパラメタ L を有しており，日本人のテキストを読む速度が一分間に 500 文字程度であることから 1CLICK-1 ではこれを $L = 500$ とした．この設定は，ユーザは 500 文字を超えたテキストは読まない，もしくは出力を読む時間が一分間しかないという状況に相当する．なお，L を無限大にすることは，ユーザはテキストをいくらでも読んでくれるという仮定を意味する．この場合，ナゲットが出力位置に応じて価値を失うことはなく，S-measure は **W-recall**[20] (重みつきナゲット再現率) に帰着する．

図 4 に，S-measure の計算例を示す．(なお，これは，S-measure にとってあまり好ましくない事例である[20]．) この例では，QA タイプのクエリ「上野動物園以外で、パンダが見られる日本の動物園は?」に対し，予め 4 つのナゲットが用意されている．各ナゲットは，ナゲット ID，重み，ナゲットの意味，および **vital strings** (ナゲットの意味を伝えるために最低限必要だと思われる文字列) の 4 つ組で表されている．(実際は，ナゲットの抽出

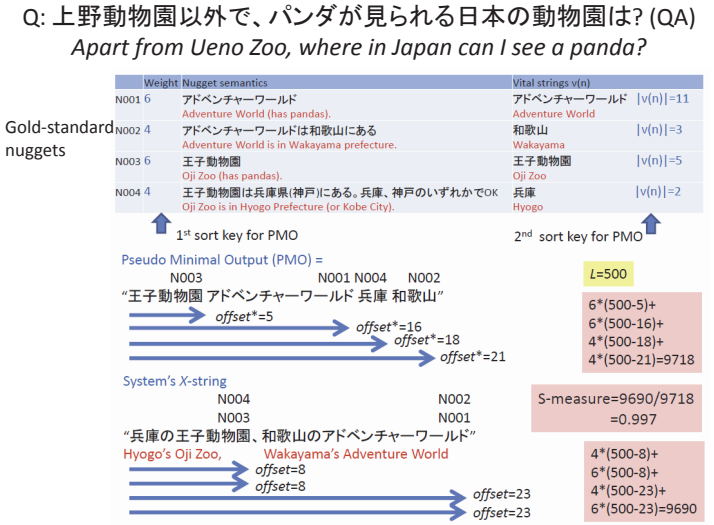

図 4 評価指標 S-measure の計算例．

元となったウェブページの URL も含めた 5 つ組である．) S-measure では，上記のようなナゲットデータに基づき，まず Pseudo Minimal Output (**PMO**: 疑似最小出力) という文字列を定義する．これはナゲット重みを第 1 キー，vital strings の長さを第 2 キーとしてナゲットをソートし，この順で vital strings を並べたもの (図の例では「王子動物園 アドベンチャーワールド 兵庫 和歌山」) であり，nDCG における正規化のための理想的な検索結果に相当するものである．この PMO に対するスコアは，図中にあるようにパラメタ $L = 500$，ナゲット重みおよび各ナゲットのテキスト先頭からのオフセット値を用いて 9718 点と算出される．一方，図中の下部にあるように，「兵庫の王子動物園、和歌山のアドベンチャーワールド」というシステム出力に対しては，そのスコアが同様に 9690 点と算出される．従ってこのシステムの S-measure は $9690/9718 = 0.997$ となる．このような新しい評価方法の利点については次節で論じる．

3.1.3 1CLICK-1 の公式結果

図 5 (左) に，1CLICK-1 の公式結果を示す．ここでは S-measure による参加システムのランキングを，W-recall によるそれと比較している．残念ながら，1CLICK-1 にはわずか 3 チーム (日本より京大，東工大，中国よりマイクロソフトリサーチアジア) しか参加がな

⋆1 この方法論は，情報検索研究の未来を議論するワークショップ SWIRL (Strategic Workshop on Information Retrieval at Lorne) II においても注目されている．http://www.cs.rmit.edu.au/swirl12/discussion.php

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

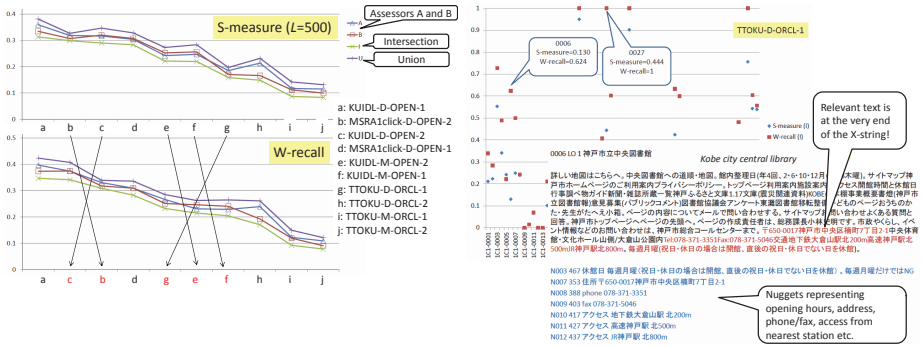

図 5 NTCIR-9 1CLICK タスクの公式結果 (左) / ある参加システムにおけるクエリ毎の S-measure と W-recall の違い (右).

かった．このうち，クエリタイプに応じた情報抽出のアプローチを採用した京大 (KUIDL)[7] が 1 位，N-gram パッセージ検索のアプローチを採用したマイクロソフトリサーチアジア (MSRA1click)[10] が 2 位のシステムを提出しており，両者の間に統計的有意差はなかった．両者とも，クエリタイプに応じて処理を切り替えており，ウェブ検索結果の他にウィキペディアとヤフー知恵袋データを活用している．一方，1CLICK タスクを検索タスクでなく，正解ナゲット抽出元のウェブページ集合から複数文書要約を生成するタスクととらえた東工大 (TTOKU)[11]⋆1は，テキストを短く収めるために，extraction ではなく abstraction (元のテキストをそのまま抜粋するのではなく独自に加工する要約技術) に挑戦しており，このようなアプローチは，限られたサイズの画面で，もしくは限られた時間内にユーザに情報を伝える 1CLICK のような状況において重要であると考えられる．

図 5 (右) は，ある参加システムについてクエリ毎に S-measure と W-recall の値を比較したグラフと，このうち両者の差が大きいクエリ 0006「神戸市立中央図書館」のシステム出力と正解ナゲットを表示したものである．図中に示したように，このシステム出力の S-measure が W-recall に比べはるかに低い理由は，このテキストの大半は不適合情報であり，適合情報はテキストの末尾の部分にやっと現れるためである．S-measure はこのように，ユーザが有用な情報に到達するまでに読まなければならないテキストの量を考慮した評価を行うことができる．なお，1CLICK-1 の Overview 論文では S-measure と readability (読みやすさ)，trustworthiness (信頼性) の関係についても考察している[21]．

⋆1 東工大のランのように正解ナゲット抽出元ウェブページ集合を既知として生成したものを，1CLICK-1 では OPEN ランと区別して **ORACLE** ランと呼んでいる．

3.2 1CLICK-2@NTCIR-10 にご参加ください!

これから 2013 年 6 月にかけて進められる第 2 回 1CLICK (1CLICK-2) では，オーガナイザに米 Northeastern 大の Virgil Pavlu らを迎え，日**英**に対象言語を拡張する．彼らは情報検索をナゲット単位で評価する新しい手法を提案しており[13]，これを応用した評価方法を導入することも視野に入れている．これに加え，1CLICK-1 とは以下の相違点がある．

- クエリタイプの細分化．1CLICK-1 で用いた 4 つのクエリタイプを細分化し，8 種類のタイプを設ける．なお 1CLICK-1 同様，要求されるナゲットの種類はクエリタイプにより異なる．
- **クエリ分類サブタスク**の導入．与えらえれたクエリを上記クエリタイプに分類する．
- Baseline ウェブ検索結果の提供．1CLICK-1 では，各自がウェブ検索 API などを用いて OPEN ランを作成する必要があったが，1CLICK-2 では，baseline 検索結果とそのウェブページの内容を参加者に提供する．参加者はこれを利用し，自前でウェブ検索を行わずにランを提出することも可能である．これは実験の再現性の観点からも有用である．
- ウィキペディアおよびヤフー知恵袋データの禁止．1CLICK-1 における DEFINITION および QA タイプの質問は，ウィキペディアおよびヤフー知恵袋データから作成されたデータであった．従って，これらを直接データベース化して検索対象とすると，問題が容易になりすぎてしまう．1CLICK-2 では上記データの直接使用を禁止し，より一般性の高い問題設定を目指す．

1CLICK-2 のスケジュールは以下のとおりである．

2012 年 4 月	(オーガナイザ) サンプルクエリ・ナゲットを公開
2012 年 8 月	(オーガナイザ) 評価用クエリを参加者に配布
2012 年 9 月	(参加者) ラン提出締切
2012 年 10 月-2013 年 1 月	(参加者・オーガナイザ・判定作業者) ナゲットマッチ判定
2013 年 2 月	(オーガナイザ) 評価結果を参加者に配布
2013 年 3 月	(参加者) 論文初稿締切
2013 年 5 月	(参加者・オーガナイザ) 論文最終稿締切
2013 年 6 月 18-21 日	NTCIR-10@NII

上記の通り，クエリセット配布が 8 月，ラン提出が 9 月と，INTENT-2 に比べ遅いスケジュールとなっている．INTENT-2 タスクで頑張った後で 1CLICK-2 タスクに取り組むことも充分可能である! 簡単な参加方法としては，前述のクエリ分類サブタスクのみへの参加

Vol.2012-IFAT-106 No.4
Vol.2012-DD-85 No.4
2012/3/26

か，baseline 検索結果もしくは正解ナゲット抽出元のウェブページからテキストを抜粋するシステムの作成が考えられる．

タスク参加申込の要領は INTENT-2 と同様である．1CLICK-1 のときと同様，参加者にはナゲットマッチ判定作業を手伝っていただく可能性が高いことにもご留意いただきたい．

4. ま と め

本稿では，NTCIR-10 の INTENT-2 および 1CLICK タスクについて紹介した．特に日本の情報検索および自然言語処理の研究者の方々にはこれらに是非参加していただき，研究の加速に役立てていただきたい．なお，NTCIR-10 向けに採択されたタスクには，INTENT-2 と 1CLICK-2 の他に CrossLink-2, PatentMT-2, RITE-2, SpokenDoc-2 およびパイロットタスクの Math がある[19]．これらのタスクへの参加もご検討いただきたい．

謝辞 NTCIR，INTENT，1CLICK 全ての関係者のご尽力に感謝する．

参 考 文 献

1) Agrawal, R., Sreenivas, G., Halverson, A. and Leong, S.: Diversifying Search Results, ACM WSDM 2009, pp.5—14 (2009).
2) Clarke, C.L., Craswell, N., Soboroff, I. and Ashkan, A.: A Comparative Analysis of Cascade Measures for Novelty and Diversity, ACM WSDM 2011, pp.75-84 (2011).
3) Clarke, C.L., Craswell, N., Soboroff, I. and Cormack, G.V.: Overview of the TREC 2010 Web Track, TREC 2010 (2011).
4) Dang, H.T. and Lin, J.: Different Structures for Evaluating Answers to Complex Questions: Pyramids Won't Topple, and Neither Will Human Assessors, ACL 2007, pp.768—775 (2007).
5) Dou, Z., Hu, S., Chen, K., Song, R. and Wen, J.-R.: Multi-dimensional search result diversification. ACM WSDM 2011, pp.475-484 (2011).
6) Han, J., Wang, Q., Orii, N., Dou, Z., Sakai, T. and Song, R.: Microsoft Research Asia at the NTCIR-9 Intent Task, NTCIR-9, pp.116-122 (2011).
7) Kato, M.P., Zhao, M., Tsukuda, K., Shoji, Y., Yamamoto, T., Ohshima, H. and Tanaka, K.: Information Extraction based Approach for the NTCIR-9 1CLICK Task, NTCIR-9, pp.202-207 (2011).
8) Li, J., Huffman, S. and Tokuda, A.: Good Abandonment in Mobile and PC Internet Search, ACM SIGIR 2009, pp.43—50 (2009).
9) Mitamura, T., Shima, H., Sakai, T., Kando, N., Mori, T., Takeda, K., Lin, C.-Y., Song, R., Lin, C.-J. and Lee, C.-W.: Overview of the NTCIR-8 ACLIA Tasks: Advanced Cross-lingual Information Access, NTCIR-8, pp.15—24 (2010).
10) Orii, N., Song, Y.-I. and Sakai, T.: Microsoft Research Asia at the NTCIR-9 1CLICK Task, NTCIR-9, pp.216-222 (2011).
11) Morita, H., Makino, T., Sakai, T., Takamura, H. and Okumura, M.: TTOKU Summarization Based Systems at NTCIR-9 1CLICK task, NTCIR-, pp.208-215 (2011).
12) Nenkova, A., Passonneau, R. and McKeown, K.: The Pyramid Method: Incorporating Human Content Selection Variation in Summarization Evaluation, ACM Transactions on Speech and Language Processing, 4(2), Article 4 (2007).
13) Pavlu, V., Rajput, S., Golbus, P.B. and Aslam, J.A.: IR System Evaluation Using Nugget-based Test Collections, ACM WSDM 2012, pp.393-402 (2012).
14) 酒井哲也: NTCIR 公式結果に基づく文書検索技術の進歩に関する一考察, FIT 2006 情報科学技術レターズ LD-007, pp.67-70 (2006).
15) 酒井哲也: よりよい検索システム実現のために: 正解の良し悪しを考慮した情報検索評価の動向, 情報処理 Vol.47, No.2, pp.147-158 (2006).
16) 酒井哲也: チュートリアル：情報検索テストコレクションと評価指標, 情報処理学会研究報告 2008-FI-89 / 2008-NL-183, pp.1-8 (2008).
17) Sakai, T.: Evaluation with Informational and Navigational Intents, ACM WWW 2012, to appear (2012).
18) Sakai, T., Joho, H.: Overview of NTCIR-9, NTCIR-9, pp.1-7 (2011).
19) 酒井哲也ほか: NTCIR-9 総括と今後の展望, 情報処理学会研究報告 2012-IFAT-106 / 2012-DD-85 (2012).
20) Sakai, T., Kato, M.P. and Song, Y.-I.: Click the Search Button and Be Happy: Evaluating Direct and Immediate Information Access, ACM CIKM 2011, pp.621-630 (2011).
21) Sakai, T., Kato, M.P. and Song, Y.-I.: Overview of NTCIR-9 1CLICK, NTCIR-9, pp.180-201 (2011).
22) Sakai, T. and Song, R.: Evaluating Diversified Search Results Using Per-Intent Graded Relevance, ACM SIGIR 2011, pp.1043-1052 (2011).
23) Santos, R.L.T., Macdonald, C. and Ounis, I.: Exploiting query reformulations for Web search result diversification. ACM WWW 2010, pp.881—890 (2010)
24) Santos, R.L.T., Macdonald, C. and Ounis, I.: University of Glasgow at the NTCIR-9 Intent task: Experiments with Terrier on Subtopic Mining and Document Ranking, NTCIR-9, pp.111-115 (2011).
25) Song, R., Zhang, M., Sakai, T., Kato, M.P., Liu, Y., Sugimoto, M., Wang, Q. and Orii, N.: Overview of the NTCIR-9 INTENT Task, NTCIR-9, pp.82-105 (2011).
26) Xue, Y., Chen, F., Zhu, T., Wnag, C., Li, Z., Liu, Y., Zhang, M., Jin, Y. and Ma, S.: THUIR at NTCIR-9 INTENT Task, NTCIR-9, pp.123-128 (2011)
27) Zhang, S., Lu, K. and Wnag, B.: ICTIR Subtopic Mining System at NTCIR-9 INTENT Task, NTCIR-9, pp.106-110 (2011)