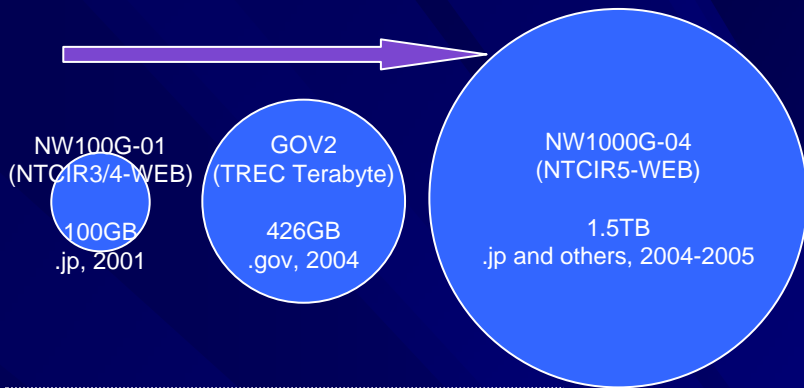


NTCIR5-WEB: テストコレクションの拡張



文書データ: NW1000G-04

- 収集対象: 日本のWebページの大半, ".jp" ドメインを中心に .jp ドメイン外も含む
- 収集機関: 2004年1月 ~ 2005年1月
- データ: 約1.5TBのテキストデータ
- サイト数: 約40万サイト
- ページ数: 約9600万ページ
- 収集ページ間のリンク数: 約9億3000万リンク

Webページの収集は早稲田大学・山名研究室との共同研究の一環として行われ、早稲田大・NIIを含む国内数拠点での並行収集を行った。

大規模WEBデータ (テラバイトデータ)



課題集合を増やす (1000課題以上)

検索課題作成方法:

課題作成者に、趣味、勉強、仕事、日常生活などに関連したできるだけ自然な検索対象事項を想起し、それに対し「代表的Webページ」をイメージして書き出してもらい、その中から既知事項検索に該当するものを選択する。検索事項は、1~3フレーズのサーチエンジンに入力すると想定される検索語で表現され、「製品・サービス」、「個人」、「名所・旧跡」、「イベント」、「オンラインショップ」など9つのカテゴリに分類される。

例:

- 「愛・地球博」(イベント) 「愛・地球博の公式ページ」
- 「トヨタ、セルシオ」(製品) 「TOYOTAセルシオのトップページ」
- 「中田英寿」(個人) 「中田英寿の公式ホームページ」
- 「楽天市場」(オンショップ) 「楽天のトップページ」

検索課題: 4セットの課題

- (1) 大学生・大学院生のアルバイト17人が作成した必須400課題
- (2) Navi-1で配布された300課題
- (3) (1)のうち比較的複雑な課題をオーガナイザが1又は2フレーズでシンプルに表現した40課題
- (4) 同じカテゴリ内で約20件ずつオーガナイザが意図的に作成したシリーズ物500課題

(1)は必須課題、(2)-(4)はオプション課題で直接評価に影響しないが、テストコレクションのさらなる強化、分析をはかる研究目的のために参加者に協力を呼びかけている

検索手法と評価

想定される検索・ランキング手法

- コンテンツベース: 個別のWebページの内容による検索方法。ページ内の文書構造を利用する方法も含む。
- ページグループ: リンクで結合された論理的なページグループを仮想ページとして利用する手法。
- アンカーテキスト: 他のWebページからのリンクの参照テキストによる検索方法。
- リンクベース: Webページの順リンク・逆リンクによる重み付けを行い、ランキングスコアを計算する手法。
- URLベース: WebページのURLアドレスの特性に基づいたランキングスコアを計算する手法。

評価指標:

適合 (検索事項に対する代表ページ)・部分適合 (代表ページを代替しうるページ)・非適合 (検索事項の代表ページに該当しない)の3段階で判定した結果をもとに、主にMRR (Mean Reciprocal Rank), DCG (Discounted Cumulative Gain) といった評価指標を用い、重複文書や密にリンクされた文書についても考慮して評価を行う予定。

- MRR: 各課題に対する実行結果リストにおける初出の正解Webページのランクの逆数を、全検索課題にわたって平均した値。正解ウェブページを上位に戻すシステムが特に有利となる。
- DCG: 正解Webページのランクのみでなく、そのページの適合度までを考慮した評価指標。

$$DCG[i] = \begin{cases} CG[i], & \text{if } i = 1 \\ DCG[i-1] + G[i]/\log_b(i), & \text{otherwise} \end{cases}$$

今後の予定

- 2005年6月: 検索結果の提出
- 2005年8月: 適合判定とシステム評価結果の返却
- 2005年12月: 成果報告会
- 2006年: テストコレクションの公開

