大規模Web情報検索実験環境の構築
Building a Large Scale Test Bed for Web Information Retrieval

大山敬三 Keizo OYAMA　高久雅生 Masao TAKAKU　相澤彰子 Akiko AIZAWA　山名早人 Hayato YAMANA

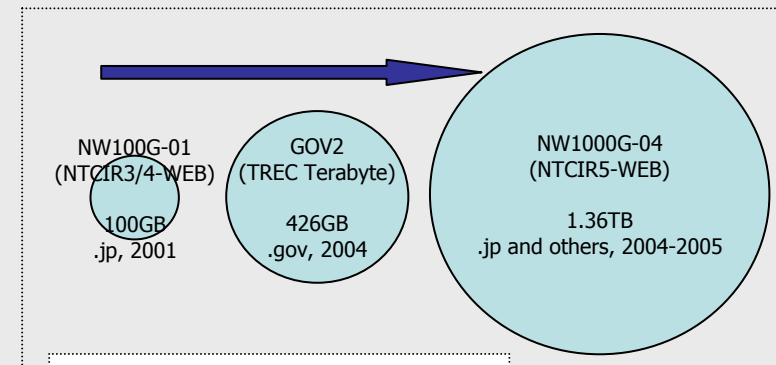
何がわかる？

「○○のページ」,「××のサイト」をピンポイントで探すサーチエンジンを作るにはどうすればいいの？

どんな研究？

たくさんの問題を用意し，いろいろな技術を使ったシステムで実験をして，検索技術を比較評価します

Web検索技術評価用テストコレクション

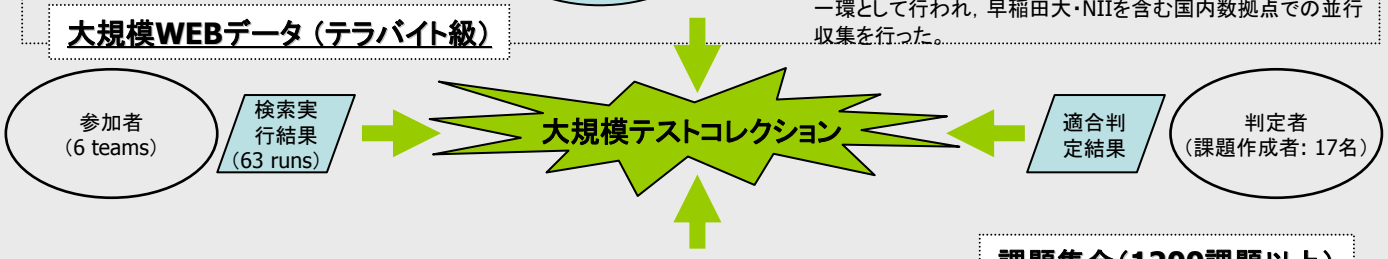


文書データ: NW1000G-04

- 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
- 収集期間：2004年1月 – 2005年1月
- Size：約1.36TB (1.5 × 10^{12} bytes) テキストデータ
- Sites：389,875 サイト
- Pages：95,870,352 ページ
- Links：1,805,160,288 リンク（収集ページ間）

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（テラバイト級）



課題集合（1200課題以上）

検索課題作成方法:

課題作成者に，趣味，勉学，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，3フレーズ以内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」(ショップ)→「楽天のトップページ」

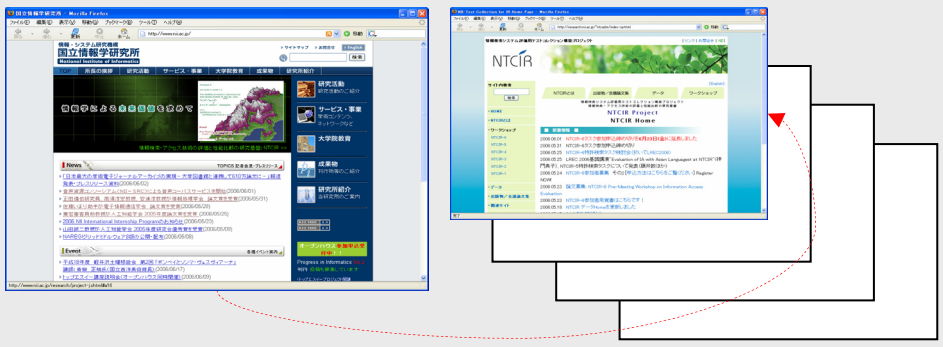
検索課題: 4セットの課題

(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41課題
(4) 同じカテゴリー内で約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た

代表的なWeb検索・ランキング手法

- **コンテント**:個別のWebページの内容による検索方法。全文(フルテキスト)を扱う。ページ内の文書構造として，タグ等の重みを加えるなども(TITLE, H1など)。
- **アンカーテキスト**:Web上のリンク部分の文字列を，リンク先の内容を表すテキストとして扱う手法。
- **リンク情報**: Webページの順リンク・逆リンクによる重み付けを行い，ページの重要度を計算する手法。インリンク数，PageRank，HITSなど。



（アンカーテキスト部分のHTMLソース）

```
... <a href="http://research.nii.ac.jp/ntcir/">情報検索評価プロジェクトNTCIR</a>
```

NII 連絡先： 大山 敬三（Keizo OYAMA）／ 国立情報学研究所 コンテンツ科学研究系 教授
TEL : 03-4212-2515　FAX : 03-3556-1916　Email : oyama@nii.ac.jp

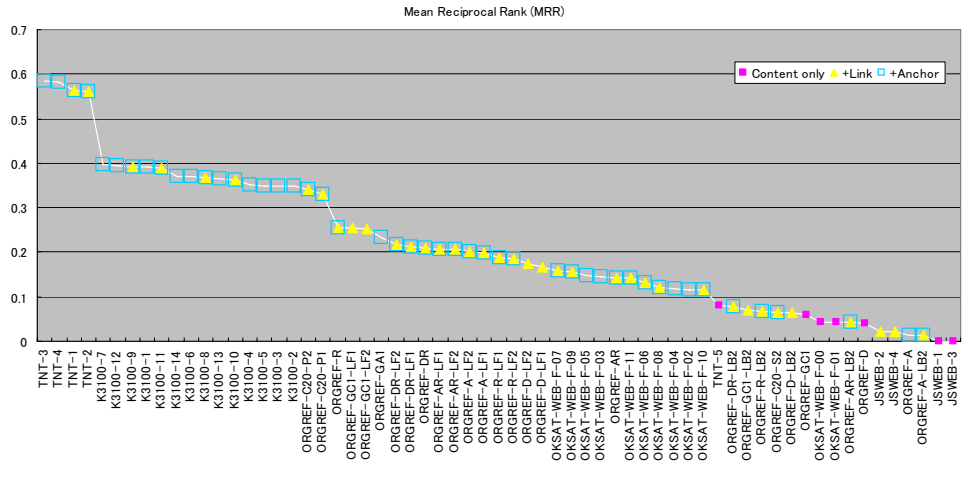
NTCIR-5 WEB ナビゲーション指向検索タスク（Navi-2）の結果

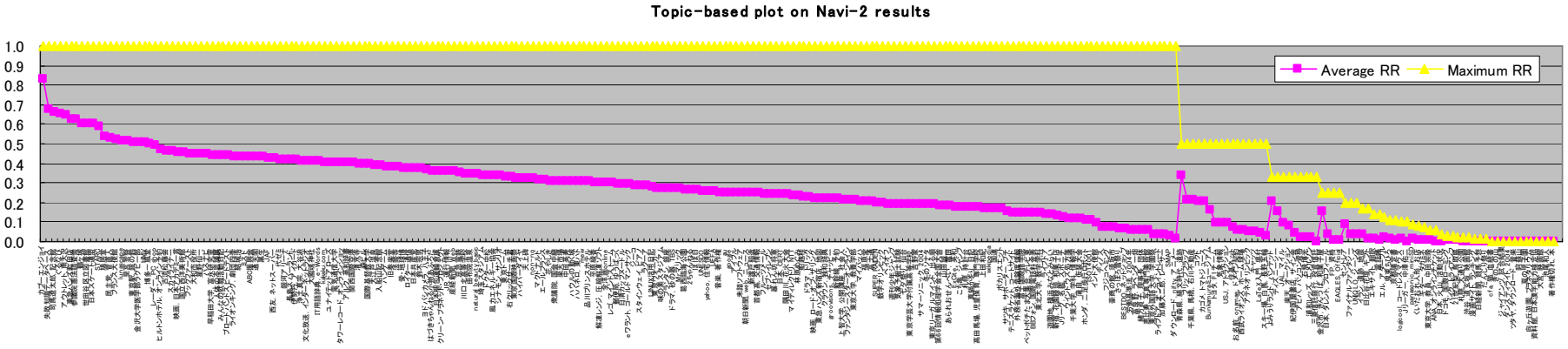
ナビゲーション指向検索： ブックマーク代わりにサーチエンジンを利用するような検索行動などとしても見られるような，特定のWebページを想定して，すぐにそのページに移れることを期待して検索をおこなうような行動。検索事項の代表的なWebページを適合とする。

検索結果としては，そのものズバリでなくとも，リンクをたどれば目的のページが簡単に見つかるような関連ページでも検索要求を満たせる。一方，誰が公開しているか分からないようなページは話題的に強い関連があっても検索意図には適さない。

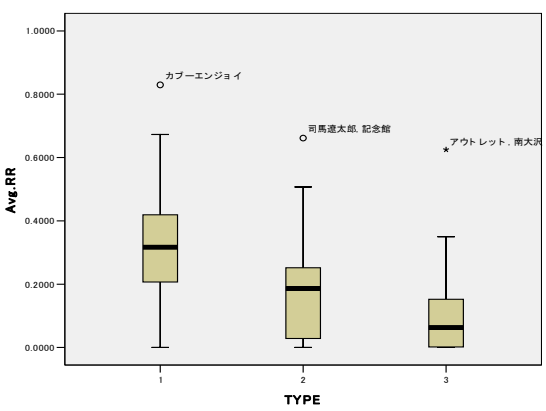
参加チームの採用した検索手法と評価結果：
高い性能を出した検索システムはアンカーテキストを効果的に使っていた。一方，リンク情報にもとづくランキング手法を採用したシステムは，それほど高い性能を出せなかった。既知事項検索には，アンカーテキストだけでも十分な性能が出ることが示唆される。

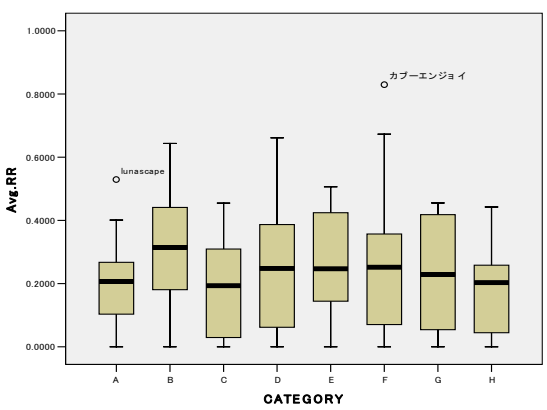
また，全システムの最高性能を見ると，大半の検索課題で正解ページを10位以内で返すことができ，ナビゲーション指向検索課題に対しては高い性能が出せることが確認できた。

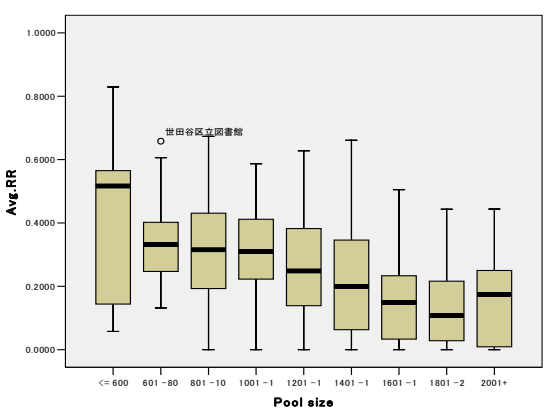




評価結果の分析







検索課題に付与したメタデータと検索難易度の関連：
TYPE（検索課題の曖昧さ），CATEGORY（検索事項の種別），SPECIALTY（検索者の対象事項への習熟度）の3点について分析を行った。TYPEについては，検索課題の特定性が高いものほど，検索システムの平均性能も高いとの結果を確認した。CATEGORYは，一部の種別間で平均性能に差異が認められた。一方，検索者の習熟度では平均性能との相関は見られなかった。

- ●（難課題種別）： 個人，製品・サービス，イベント
- ●（中程度）： オンラインショップ・サービス，情報資源，施設，名所・公園
- ●（易課題種別）： 企業・組織

検索難易度の推定： 複数の検索システムの結果である，プールサイズと平均性能との間に相関関係が見られた。今後，これらの情報を用い，検索式の情報＋αから，あらかじめその検索結果の適合度を示すなどの展開が期待できる。

今後の展開

ナビゲーション指向検索にはアンカーテキストが効果的であることがわかったが，一方で新しく公開されたばかりのページなどではアンカーテキストは利用できない。今後はアンカーテキストが有効かどうかを判別する方法と，有効でない場合に適した検索技術を検討・評価することが必要である。

NII 連絡先： 大山 敬三（Keizo OYAMA）／ 国立情報学研究所 コンテンツ科学研究系 教授
TEL：03-4212-2515 FAX：03-3556-1916 Email：oyama@nii.ac.jp