データベースシステム 100−16
（1994. 10. 14）

日本語情報検索システムのためのベンチマークの構築

小川泰嗣（リコー）、木本晴夫、田中智博（NTT）
石川徹也、増永良文（図書館情報大）、芥子育雄（シャープ）
豊浦潤（RWCP）、福島俊一（日本電気）、宮内忠信（富士ゼロックス）

日本語を対象とした情報検索に関する研究開発では、性能（検索精度）の評価に開発元独自の方法が用いられてきた。このような状況に対し、われわれは「情報検索システム評価用データベース構築ワーキンググループ」を設置し、情報検索手法・システムを公正かつ客観的に評価するためのベンチマーク構築を目指している。本稿では、情報検索システムモデル・対象データの特性とサンプル件数・評価法・作成手順などベンチマーク構築に関する現在までの検討内容を報告する。

Development of a Benchmark for Japanese Information Retrieval Systems

OGAWA Yasushi (Ricoh), KIMOTO Haruo, TANAKA Tomohiro (NTT),
ISHIKAWA Tetsuya, MASUNAGA Yoshifumi (ULIS), KESHI Ikuo (Sharp),
FUKUSHIMA Toshikazu (NEC), TOYOURA Jun (RWCP),
MIYAUCHI Tadanobu (Fuji Xerox)

In the research and development of Japanese information retrieval systems, different research groups have been using different measures to evaluate their system because there is no standard or *benchmark* for them. Our working group in IPSJ-SIGDBS has been developing such a benchmark, and in this report we will present several of its characteristics: IR models, the features and size of texts and queries, evaluation methods, and the development procedure of the benchmark.

1 はじめに

研究開発においては新たな手法・システムの有効性を示すために公平かつ客観的な評価が不可欠であり、評価のための共通の基盤であるベンチマークの存在は極めて重要である。テキストデータベース・自動索引などを含めた情報検索も例外ではなく、様々な研究開発が行なわれるにしたがって、ベンチマークの構築が行なわれてきている。欧米では、CACM や Medlars など主に論文を対象としたベンチマーク（テストコレクション）が作成され、一般に提供されてきている [1]。また、1992 年からはアメリカ規格協会（NIST）が大規模テキスト DB を対象としたテキスト検索システムの評価会 (TREC: Text REtrieval Conference) を行なっている [4][5]。

ベンチマークの評価項目として一般的には処理速度と処理性能の二つの面が考えられるが、これらベンチマークでは処理精度（検索精度）のみが取り上げられており、データベースのベンチマーク [10] では処理速度が評価項目となっているのと対照的である。これは以下の理由による。データベースにおいては、データモデルにより処理内容が明確に規定されているため処理精度は 100% が前提となっており、処理速度に対する評価のみが意味を持つ。一方、情報検索においては、処理対象が文書という自然言語であるため解釈・理解といったあいまい性を含むこととなり、処理速度よりも処理精度に対する関心が集中している[1]。

日本においても古くから情報検索の研究開発 [3][11] が行なわれてきたが、評価は研究開発元が独自に準備したデータ・評価方法に基づいて行われてきた。これは、前述のように情報検索では対象が自然言語なので手法やパラメータ設定が言語依存であり、欧米で作成された英語用ベンチマークが適用不可能だからである。その一方で、電子化データの蓄積の遅れ、文化的な相違などの理由から、日本語用のベンチマークが存在していないためである。その結果、各手法の有効性の検証、手法間の優劣の判定が公正かつ客観的に行なわれてきたとは言えない [8]。したがって、日本における情報検索の研究開発の一層の促進のために、日本語ベンチマークの構築が火急の課題となっている [6]。

そこで、われわれは日本語文書を対象とした情報検索システム用ベンチマークの作成を目的とし、情報処理学会データベースシステム研究会の下部組織として「情報検索システム評価用データベース構築ワーキンググループ」を 1993 年 2 月に設立、活動を続けている。これまでに、18 回の会合を持ち、ベンチマーク作成に関する様々な検討を行なってきた [7][9]。

現在までに、ベンチマークで扱う情報検索システムモデルの明確化、対象文書・質問文の特性とサンプル件数、システム評価方法について論議し、まとめの段階にある。ベンチマーク作成は二段階で行なうことにしている。すなわち、これまでの議論に基づいてテスト版を作成し、多くの研究開発者に配布、意見を募る。それに基づいて改良を行ない、データ件数などの拡充を計った完成版を作成する予定である。現在、テスト版作成の前準備を行なっている。

以下、ワーキンググループ（以下 WG）での検討内容を紹介する。まず、2 章で情報検索システム用のベンチマーク構築において考慮すべき項目について検討する。3 章で各検討項目に対する本 WG の方針を示した後、ベンチマーク作成手順について 4 章で述べる。最後の 5 章はまとめで、本ベンチマークと既存のものとの比較を表形式で示す。

2 ベンチマーク構築上の検討項目

ベンチマークは、その目的に応じて、対象とするシステム・データ・評価手法等が異なる。この章では、ベンチマーク構築にあたって検討すべき項目を上げる。

2.1 情報検索システムモデル

情報検索システムには、様々な形態が考えられる。したがって、ベンチマーク作成にあたっては以下の点を検討し、モデルを明確化しておく必要がある。

[1] だからといって、処理速度が無視されているわけではもちろんない。

- 検索タイプ
  TREC では、大量の文書データベースからユーザが発行する検索要求に該当する文書を得る adhoc と、ユーザがあらかじめ登録しておいた要求に該当する文書を新たに発生した文書から選択する routing (filtering) に分類している。
- 検索単位
  検索要求に対して、文書（識別子）を解答とするのか、文書のなかの特定部分を選出して解答とするか。
- ランク付け
  解答を、検索単位の集合とするのか、評価を計算しランク付けするのか。
- 対話機能
  必要な情報を得るための対話機能を保有しているか否か。

2.2 対象文書

検索システムが扱う文書の特性として以下の項目がある。

- 種類
  新聞記事・特許・論文など様々なものがある。
- 分野
  新聞では政治・経済・社会・スポーツ等、論文では物理・医学・経済等の分野が考えられる。分野を特定するか、様々な分野を扱うのか。
- タイプ
  例えば、新聞では報道記事、論説記事、株価の一覧など多様である。
- 書誌項目
- 図表など
  本文以外に、図表・写真などを含めるか否か。

2.3 質問文

対象文書に応じて質問文の特性も変わってくる。その他、質問文の機能として捉えた場合、どのような処理が必要かも様々である。

- 形態
  キーワード、Boolean 演算子、自然言語文などがある。
- 用語の制限
  キーワード方式の場合、統制語・自由語いずれの方式とするか。
- 付加項目
  TREC では、検索要求文のほかに詳細記述・キーワードなどの補助項目も付加している。このような情報を用いるか。

2.4 規模

情報検索ベンチマークでは、現実の環境における検索精度の測定は作業量的に極めて困難であるため、サンプリングにより実作業可能な規模で行なうことを前提としている。したがって、対象文書・質問文に関連して サンプル数の設定が重要である。

ベンチマークの有効性・信頼性からはサンプル文書数は多い方が良い。（有効性に関しては、一般的に高度な処理を用いることで性能向上をはかることができるが、そのために必要な辞書等のデータ作成作業が増加するため、高度な手法が実用規模で有効かを示すにはある程度のサンプル数が必要である。信頼性に関しては、ベンチマークが現実世界のサンプルであるため、統計学的に意味のある規模でなければならない。）一方、ベンチマークの効率化の観点からはサンプル文書数は少ない方がよい。

- サンプル対象文書数
- サンプル質問文数

2.5 評価法

情報検索ベンチマークは処理精度の評価を目的としており、以下の検討項目がある。

- 評価指標
  処理精度の評価値として、次式のように定義される再現率（検索洩れの少なさの指標）・適合率（検索ノイズの少なさの指標）が一般

的に用いられている。

$$再現率 = \frac{検索文書集合中の正解文書数}{正解文書数} \quad (1)$$

$$適合率 = \frac{検索文書集合中の正解文書数}{検索文書数} \quad (2)$$

ランキング機能を有する検索システムに対する再現率・適合率の計算法も標準化しなければならない。

また、再現率・適合率以外の指標を導入するか否かも検討しなければならない。例えば、対話を想定している場合、正解を得るのに要する対話回数等の指標が必要になる。

- 正解判定単位
  再現率・適合率の計算では、質問文に対する正解文書集合が定義されていなければならないが、正解の決定は検索モデル依存なので注意が必要である。例えば、文書の該当部分を検索結果とするモデルでは、文書として正解であっても、質問文に該当する部分を正しく抽出していなければ正解と判定できない。
- 正解レベル
  質問文に対し各文書を正解・不正解の二値で判断するか、レベル分けするか。
- 検定法
  ある手法の有効性を検証するには、指標値（の平均値）を単純に比較するだけでは不十分であり、統計的に有意な差があることを示す必要がある。したがって、ベンチマークとして検定法を規定する必要がある。

2.6 その他

これまで上げた項目以外にも、ベンチマークとして成立するためには以下の項目も検討する必要がある。

- 配布媒体
  フロッピー、CD-ROM、テープ、ネットワーク経由などがある。
- データフォーマット
  ベンチマークは様々なデータから構成される。その論理データフォーマットを規定しなければならない。
- 著作権・使用料
  対象文書・質問文等の著作権の所在を明確化しておく必要がある。ベンチマークとして広く利用されるためには、著作権使用料は無料または極めて安価でなければならない。
- 監査
  正式には第三者による監査が実施されることが望ましいが、研究目的のベンチマークでは監査を導入していないものも多い。

3 本WGの方針

前節で述べたような項目に対し、本WGとしては以下のように考えた。

3.1 情報検索システムモデル

情報検索の基本機能の評価をベンチマークの対象とし、単純なモデルを想定した。

- 検索タイプ：adhoc
- 検索単位：文書
- ランク付け：なし[2]
- 対話機能：なし

ここで想定した情報検索システムモデルを図示すると図 1 のようになる。このモデルを前提とすると、対象文書および質問文の特性と規模、正解集合を含む評価法を定義することで、ベンチマークを構築できる。（ベンチマークとしては、評価対象システムの手法・実装法は問わず、システムをブラックボックスとして扱う。）

[2]評価対象のシステムがランク付け機能を持っていてもかまわないが、本ベンチマークとしてはランク付け機能に対する有効性の評価法を規定しない。対話機能についても同様。

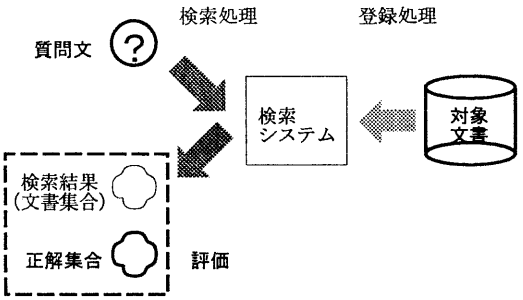

図 1: 情報検索システムモデル

3.2 対象文書

- 種類：新聞記事
- 分野：経済

  情報検索のニーズは、新聞の記事検索、特に経済分野が多いとの検討結果に基づき決定した。

  経済分野の判定は経済面に記述されているかで判定する予定。
- タイプ：特定せず
- 書誌項目：含める

  発行日、朝夕刊の区別、紙面等を書誌項目とする。
- 図表など：含めない

  対象は記事のタイトルおよび本文とし、図表・写真は含めない。

3.3 質問文

広い範囲の検索モデルへの対応を考え、質問文の形態は自然言語文、付加項目はなしとした。自然言語文として「～に関する記事が欲しい」という形式とする。

既発表 [7] で述べたように、質問文はその質問文を正しく処理するために要求される機能によって分類できる。現在は、以前の分類を整理しつぎの 5 つの機能に分類した。

1. 基本機能

   質問文は単語あるは単純な名詞句のみで構成され、文書に単語あるいはそれをシソーラスで展開したものが含まれているか否かだけで、正解が決定できるもの。

   例 1：パテント…
   （「パテント」の同義語「特許」）

   例 2：自動車の軽量化技術…
   （「自動車」全体に対する「エンジン」部分）

2. レンジ機能

   数値などの範囲を正しく処理できるもの。数値の大小比較、単位の理解・変換などの機能が必要になる。

   例：一千万円以上の脱税詐欺事件…

3. 構文解析機能

   質問文に、動作およびその主体・対象が記述されているもの。質問文（および検索対象文書）を構文解析する等の処理が必要になる。

   例：政府の鉄鋼業界に対する支援策…

4. 内容解析機能

   質問文だけでなく、検索対象文書の内容解析が行なわれないと正しく処理できないもの。対象文書の文脈解析等の処理が必要になる。

   例:不況によるコンピュータ業界の雇用調整…

5. 知識処理機能

   単語の展開、文の解析だけなく、知識を用いた推論などを用いなければ正しく処理できないもの。

   例：戦後の円の最高値…
   （戦後に関する知識が必要）

3.4 規模

ベンチマークの環境を統計学的にモデル化することで、サンプル件数に応じた誤差を推定することができる [9]。この理論にしたがい、作業量なども考慮して、ベンチマークの規模（サンプル数）を定めた。

- 対象文書数：6000（完成版）、600（テスト版）

  対象文書数ηに関しては、検索条件に対する正解文書集合の最小文書数λ_{min}・最大文書数λ_{max}、および全文書集合に対する正解文書集合の割合の最大許容値c_{max}から決めるべきであるとの結論を得ている。

  完成版においては、$\lambda_{min} = 30$（統計上の最低数）、$\lambda_{max} = 120$（再現率を信頼区間95% 許容誤差 10%以下で推定できる値）、$c_{max} = 0.02$とし、対象文書数ηを6000件程度とする。ただし、これは現時点での計画であり、テスト版に対する意見や作業量などを考慮して最終決定する。

  テスト版においては、統計学上の議論よりも作業量を優先し、WG メンバーで対応可能である600件で行なう。（この時、$\lambda_{min} = 5$、$\lambda_{max} = 30$、$c_{max} = 0.02$である。）

- 質問文数：機能ごと30、合計150

  質問文については、前述の機能ごとに統計上の最低数の30件ずつ用意することとする（テスト版、完成版とも）。したがって、質問文の機能を5つ想定しているので、合計150件となる。

3.5 評価項目（正解集合）

評価項目に関しては、以下のように定めた。

- 評価指標：再現率・適合率

  ランキング機能を有する検索システムに対する再現率・適合率の計算法の規定は検討中。

- 正解判定単位：文書

- 正解レベル：多段階

  WG 内での正解集合作成の試行テストにより、正解とも不正解とも判断しがたい場合がかなり存在することがわかった。既存のベンチマークでも正解にレベルを設定しているものあるので、正解・不正解に二分するのではなく、判定を多段階とし、あいまいな文書の存在を許容することとした。段階としては、cranfield のもの[2]を参考に、例えば(1) 質問文の内容を主題とした文書、(2) 質問文の内容を少しでも記述している文書、(3) 質問文の内容を直接は記述していないが関連した内容を記述している記事、のようなものを想定しているが、詳細は検討中である。

- 検定法：符合検定[12]

  再現率・適合率の分布がわからないため、値の分布がわからない場合に適用可能な検定法である符合検定を採用した。

再現率・適合率を計算できるように、質問文ごとの正解文書集合を提供する。さらに、正解データとして、正解の判定理由も添付することにしている。これは、質問文は一文という制限を設け、TREC におけるような詳細記述を添付しないため、正解判定にあいまいさが生じるためである。

なお、2.5 節であげていなかったが WG での試行テストから明らかになった問題点として、判定者によるバラツキの大きさがある。すなわち、同一の対象文書集合と質問文に対しても、正解文書集合が判定する人によってかなり異なったのである。この問題に関しては、質問文について二人の正解判定者を割り当て、両者の判定が食い違った場合は協議によって正解を決定することとし、判定者によるバラツキを小さくする。

3.6 その他

- 配布媒体

  テスト版では、件数が少ないので作成も簡単なフロッピーディスクによる配布を考えている。一方、最終版は、件数の多さなどからCD-ROM を予定している。

- データフォーマット

  一文書（記事）は本文の他にも書誌項目を含むため、文書の記述に SGML の利用を検討している。ただし、詳細は未定である。

- 著作権・使用料

  対象文書の著作権はデータ提供者、質問文・正解集合は本 WG にある。対象文書の著作権使用料は、できるだけ安価なものとするべくデータ提供者と交渉中である。

- 監査方法
  今回は研究システムの評価を第一目的としており、費用の点でも監査機関の設置は不可能である。そこで、特に監査は定めず、各研究者・開発者の良識に任せることとする。

4 ベンチマーク作成手順

前章で説明した方針に基づきベンチマーク作成手順を示す。ここで示すのはテスト版作成の手順であり、作業者は WG メンバー（10 名：現時点は 9 名であるが 1 名増加の予定）である。作業手順のポイントは、作業者が複数居ることを考慮して、各作業者は対象文書の一部のみを担当する点である。これは、テスト版では文書数を 600 と少なくしているが、それでも人手で 600 件の記事をすべて見るのは大変な作業だからである。

1. 母文書集合の用意

   某全国紙、朝刊、経済面 4 ヶ月分（93/9/1 ～ 93/12/31）とする。

   母集合の件数は、現在入手手続き中であり正確な値はわからないが、約 6000 件と推定される。各記事に日付、紙面の順に一意の番号を付与する。

2. 対象文書集合の作成

   母文書集合から 600 件を無作為にサンプリングし、対象文書集合とする。

3. 対象文書集合の割り当て

   600 文書を各作業者に割り当てる。一文書に二名の作業者を割り当てるので、各作業者の分担は $2 * 600/10 = 120$ である。

4. 質問文の作成

   割り当て記事を参考に、各機能ごと 3 個、計 15 個の質問文を作成する。（全作業者の質問文を集めると、各機能ごと 30 個、計 150 個となる。）なお、作業者間で重複がないように調整も行なう。

5. 質問文作成者による正解文書集合の作成

   質問文作成者は、自分の作成した質問文 15 個について自分の割り当て文書 120 件に関し正解判定を行なう。その際、判定ランクと判定理由も記述する。

6. 質問文作成者以外による正解文書集合の作成

   各作業者は、自分以外の作成した質問文 $15 * 9 = 135$ 個について自分の割り当て文書に関し正解判定を行なう。その際、各質問文に添付されている判定理由を参考にする。

7. 正解文書の調整

   同一文書を割り当てられている二人の作業者は、正解文書のつき合わせを行ない、判定が割れたものについては調整を行なう。

8. 質問文と正解文書の確定

   全ての作業者からの正解文書を集める。各質問文について、正解文書数が $[5(=\lambda_{min}), 30(=\lambda_{max})]$ の範囲になっているか検査する。条件を満たしていれば、その質問文と正解集合をベンチマークに用いるものとして確定する。条件を満たしていない場合、その質問文はベンチマークに含めることはできないので、新たな質問文を作成し、ステップ (5) 以降の作業を繰り返す。

   ベンチマーク用の質問文と正解集合の組が各機能ごと 30 個、計 150 個となったら終了。

5 おわりに

日本語情報検索システムのためのベンチマーク構築を目的とした WG における検討内容を報告した。現在は、残課題の検討を進めている。新聞社との契約が終了していないため、テスト版作成には取り掛かっていないが、契約後直ちに実作業に入り、今年度中に配布したいと考えている。

最後に、既存ベンチマークの CACM・TREC-2 と本ベンチマーク（JIRB：仮称）の比較を表 1 に示す。本ベンチマークは未完成なので、予定値を示す（"[ ]" 内はテスト版）。質問文の数や多様さ、正解レベルの多段階設定などの点で、他ベンチマークより JIRB では詳細な性能評価が可能である。また、対象文書数が極めて多い TREC-2 でも評価に用いられる文書数は少なく、JIRB は規模の点でもひけを取らない。

表 1: 既存ベンチマークとの比較

	CACM	TREC-2(adhoc)	JIRB（仮称）
作成	Cornel Univ. (E.Fox)	NIST	情報処理学会 DBS 研
発行年	1983	1992	1996 [1995.3]
媒体	CD-ROM	コンテスト データのみ CD-ROM で 入手可能	CD-ROM [フロッピー]
対象	論文	新聞・論文等	新聞
分野	計算機科学	不特定	経済
文書数	3,204	1,078,925 評価における 1 質問文 当りの文書数は 1,106	6,000 [600]
質問数	52	50	150（機能ごと 30）
質問文	自然言語文, Boolean	自然言語文 キーワード等の付加項目	自然言語文
正解判定	正解・不正解	正解・不正解	多段階

本ベンチマークおよび WG の活動に関しコメント・意見のある方は WG リーダ木本まで御連絡下さい。(Tel:0468-59-3387, Fax: 0468-55-1152, e-mail: *kimoto@syrinx.ntt.jp*)

参考文献

[1] E.A. Fox. Characterization of two new experimental collections in computer and information science containing textual bibliographic concepts. Technical Report 83-561, Cornel Univ., 1983.

[2] E.A. Fox, editor. *VIRGINIA DISC ONE* (CD-ROM). Virginia Poly. Univ., 1990.

[3] 藤澤 浩道, 絹川 博之. 情報検索における自然言語処理. 情報処理, Vol. 34, No. 10, pp. 1259–1265, 1993.

[4] D. Harman, editor. *The 1st Text REtrieval Conference (TREC-1).* National Institute of Standards and Technology, 1992.

[5] D. Harman, editor. *The 2st Text REtrieval Conference (TREC-2).* National Institute of Standards and Technology, 1993.

[6] 石川 徹也ほか. 自動索引システム評価用ベンチマークテキストＤＢの構築. 情報処理学会研究会報告, Vol. DBS90, pp. 93–95, 1992.

[7] 石川 徹也ほか. 情報検索システムの評価のためのベンチマークデータベースの構築. アドバンストデータベースシステムシンポジウム *93*, pp. 217–226, 1993.

[8] 木本 晴夫. 自動索引システムと情報検索システムの評価用共通データベースの事例. 情報処理学会研究会報告, Vol. DBS90, pp. 83–92, 1992.

[9] 木本 晴夫ほか. 情報検索システムの評価用データベースの構築の提案. 情報処理学会研究会報告, Vol. FI20, No. 1, pp. 1–8, 1993.

[10] 喜連川 優ほか. データベース処理におけるベンチマーク. 情報処理, Vol. 31, No. 3, pp. 328–342, 1990.

[11] 小川 泰嗣. テキストデータベース技術の最近の動向. アドバンストデータベースシステムシンポジウム *93*, pp. 153–162, 1993.

[12] 応用統計ハンドブック編集委員会. 応用統計ハンドブック. 養賢堂, 1989.