情報検索システム評価用ベンチマークVer.1.0（ＢＭＩＲ－Ｊ１）について

芥子育雄（シャープ）、木本晴夫、田中智博（ＮＴＴ）、石川徹也、
増永良文（図書館情報大）、小川泰嗣（リコー）、豊浦 潤（ＲＷＣＰ）、
福島俊一（日本電気）、宮内忠信（富士ゼロックス）、三池誠司（東芝）、
松井くにお（富士通研究所）、木谷 強 （ＮＴＴデータ通信）

1996年3月にモニター公開を予定している，日本語テキストを対象とした，初めての情報検索システム評価用ベンチマークについて報告する．本ベンチマークは，600件の対象文書，60件の検索要求，及び正解集合（検索要求に対する最大正解文書数30件，最小正解文書数5件）から構成される．本ベンチマークの特徴は，情報検索システムに求められるシステム機能を5種類に分類し，各検索要求を正しく処理して正解を検索するために必要なシステム機能を付与した点にある．このため，各種の検索手法の優位性判定に本ベンチマークを利用することが可能である．

Overview of BenchMark for Japanese IR System Ver. 1.0 (BMIR-J1)

KESHI Ikuo (Sharp), KIMOTO Haruo, TANAKA Tomohiro (NTT),
ISHIKAWA Tetsuya, MASUNAGA Yoshifumi (ULIS),
OGAWA Yasushi (Ricoh), TOYOURA Jun (RWCP),
FUKUSHIMA Toshikazu (NEC), MIYAUCHI Tadanobu (Fuji Xerox),
MIIKE Seiji (Toshiba), MATSUI Kunio (Fujitsu Lab.),
KITANI Tsuyoshi (NTT Data)

In this paper, the first benchmark for Japanese information retrieval system is presented. It consists of three basic parts as follows: 600 documents, 60 questions, and right answers, which are from 5 to 30 documents, given to each question. Our working group in IPSJ-SIGDBS has developed such a benchmark, and proposed 5 functions which the best IR system should have and marked functions which were needed to get right answers to each question. By this major characteristics, the benchmark can be applied to evaluate various kinds of retrieval methods.

1．はじめに

近年，インターネットの急速な発展に伴って，新聞記事，論文，特許など，利用者にとって極めて有用な情報がコンピュータネットワークから容易に入手できるようになってきており，これらテキスト情報の洪水の中から適切な情報を見付け出す検索システムに対する必要性が高まっている。米国では，１９９２年から，テキスト検索システムの研究開発の活性化を目的に，大規模かつ分野を限定しない対象文書（百万文書，約2Gバイト）を元にした評価会TREC-1[1], TREC-2[2], TREC-3[3], TREC-4[4]が開催されている。TRECに参加した検索システムの数は，TREC-1の２４システムに対し，TREC-4では，３６システムと増加しており，TRECの目的は達成されている．

検索システムの性能を客観的に比較・評価するためには，TREC のような共通のベンチマーク（テストコレクション）の存在は極めて重要である．実際，欧米ではTREC以前からCACMやMEDLARSなど主に論文のアブストラクトを対象にした小規模なテストコレクションが作成され，広く提供されている[5], [7]．特に，論文発表において，新たな手法の優位性を示すために，これらのテストコレクションが利用されてきた．しかし，今日まで，日本語テキストを対象とした検索システム評価のための共通のベンチマークは存在しなかった．このため，開発元独自に準備したデータに基づいて評価が行われており，客観性に欠ける問題があった[6]．

そこで，我々は日本語テキストを対象とした情報検索システム評価用ベンチマークの作成を目的に，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月に設立し，ベンチマーク作成に関するさまざまな検討を行ってきた．その成果として，１９９６年３月に「情報検索システム評価用ベンチマーク　Ver. 1.0」(BenchMark for Japanese IR System Ver. 1.0以下，BMIR-J1)を，テスト版の形でモニター公開する予定である．情報検索システムの評価項目には，検索効率（検索速度・インデックスオーバーヘッド）と検索効果（検索要求と検索結果の関連度）がある．しかし，情報検索システムにとっては，検索効率が良くても，検索効果が悪ければその価値はないので，我々は検索効果の評価に焦点をあてた．

我々は，これまでに，ベンチマークの主要構成要素である対象文書の規模，検索要求の意味，正解集合の作成方法について主に検討を行ってきた[8], [9], [10]．検索効果を評価するために，一般に，適合率と再現率が用いられる．適合率は検索システムによって検索された文書中での正解文書の割合，再現率は全正解文書中での検索された正解文書の割合であり，その信頼性は対象文書の規模に依存すると考えられる．このためTRECでは百万文書を対象にしたテストコレクションが作成されている．しかし，百万文書すべてを対象に正解判定を行うことは不可能であり，TRECでは参加した検索システムの上位の検索結果の集合を正解判定の対象としている．これでは，検索要求を現状のシステムが正しく処理できる範囲を対象に正解集合を作成していることになり，実際の全文書を対象にした適合率，再現率を推定することはできない．MEDLARS, CACMなどではサンプル文書を対象に正解集合を求めているが，対象文書の規模についての議論はなされていない．我々は，適合率，再現率を統計学的に推定できる方法を理論化し，この理論に従い，ベンチマークの対象文書数を６０００件，各検索要求に対する正解集合の最小文書数を３０件，最大文書数を１２０件とした[8]．テスト版のベンチマーク BMIR-J1では，作業量を考慮して，対象文書数を６００件，正解集合の最小文書数を５件，最大文書数を３０件とした．

ユーザによる実際の検索要求は，多岐に渡る．我々は，これらの検索要求を正しく処理するために必要なシステム機能を分類整理した後，それらのシステム機能を評価する観点から検索要求を作成した[9]．この結果，BMIR-J1では，６０件の検索要求を用意した．

BMIR-J1 作成の過程で，６００件の対象文書について，６０件の検索要求の正解集合を求める作業が最も困難である．我々は，その作業手順も明らかにしている．BMIR-J1の作成作業は，本ワーキンググループのメンバー１２名で全て行ない，今後の本格版作成への方法論を確立した[１０]．

２．BMIR-J1の構成要素

BMIR-J1 は，欧米における情報検索システム評価用ベンチマークと同様に，対象文書，検索要求，正解集合から構成される．

２．１　対象文書

文書の母集合は，日本経済新聞社の経済面４ヵ月分（９３年９月１日～９３年１２月３１日）から，検索対象として適切でない「人事異動」や「お断り」などの記事を除いた，約４万件である．この母集合から，６００件を無作為にサンプリングし対象文書とした．

対象文書は，SGMLを利用した情報検索システムの評価にも適用できるように，以下に示す通り，SGML形式に変換した．各記事は，記事ID，ヘッダ（見出し，書誌事項），本文（段落）から構成されている．BMIR-J1には，この他，日本経済新聞社から提供された各記事のキーワードも付与されている．

<記事>
<記事ID>09010003</記事ID>
<ヘッダ>
<見出し>ＮＴＴ、希望退職１万人募集。</見出し>
<発行年月日>93年9月1日</発行年月日>
<媒体>日本経済新聞　朝刊</媒体>
<紙面>1</紙面>
</ヘッダ>
<本文>
<段落>
日本電信電話（ＮＴＴ）は三十一日、経営合理化の一環として十月一日から一般社員を対象に約一万人の希望退職者を募集すると発表した。対象者は満四十歳以上五十七歳以下の勤続十年以上の社員で、規定の退職金に加え基本給の九カ月または十二カ月分の「特別一時金」を支給する。すでに同社の組合である全電通とは実施について合意している。同社は管理職についても希望退職者の募集を検討している。（解説１０面に）</段落>
・
・
</本文>
</記事>

２．２　検索要求

検索要求としては，キーワードの論理演算，自然言語文（１文），トピック（複数の文）などが考えられるが，検索要求は自然言語文形式で表現するのが，ユーザにとっては簡単かつ自然なので，BMIR-J1 では，検索要求として１文の自然言語文に制限した．キーワードの論理演算を Query（検索システ

ムにおける検索要求の表現形式）とする情報検索システムでは，検索要求を元にマニュアルでQueryを作成することで，本ベンチマークを利用できる．TRECでも，これまではマニュアルによるQuery作成を考慮して，トピック形式で検索要求が与えられていたが，実際のユーザ要求を反映して，TREC-4からは１文の自然言語文で検索要求が与えられている[４]．

新聞記事に対する検索要求は，一般に，「～を主題とする記事が欲しい」，「～に関連した記事が欲しい」のような形式が取られる．BMIR-J1では，検索要求として，検索内容を指定する「～」の箇所のみを採用した．「主題とする」や「関連した」のような検索結果と検索要求の類似度に関する部分の評価は，次節でのべる正解集合のレベル分けによって行う．

BMIR-J1の検索要求は，付録に示す６０件である．検索要求は，実際のユーザ要求を反映したものとなっていることが理想的である．しかし，実際の検索要求を数多く収集して分析するすることは，作業量の点で困難である．BMIR-J1では，情報検索システムに要求されるシステム機能を抽出・整理することによって，そのシステム機能を評価する目的で，トップダウンに検索要求を作成するという手法を取った．BMIR-J1では，次の５種類のシステム機能に分類した[１０]．

・Ｆ１：基本機能
　検索要求は，単語あるいは単純な名詞句のみで構成され，それらの単語あるいは，それらをシソーラス等により展開したものの論理演算（AND, OR）により正解が決定できるもの．
　例：「国内航空大手３社」（「全日空」，「日本航空」，「日本エアシステム」への展開）
・Ｆ２：数値レンジ機能
　検索要求は，数値を含み，正解を決定するためには，数値の単位の認識や大小比較を必要とするもの．例：「１ドル＝１０５円以上の円高」
・Ｆ３：構文解析機能
　検索要求に，動作およびその主体・対象が記述されているもの．例：「コンピュータメーカーの人員削減」（動作：削減，主体：コンピュータメーカー，対象：人員）
・Ｆ４：内容解析機能
　検索要求に，深い言語知識を必要とする表現が含まれており，正解を決定するためには，対象文書の内容解析を必要とするもの．例：「株価動向」（動向表現の理解，対象文書中で株価の動きが記述されているかどうかの判定が必要）
・Ｆ５：知識処理機能
　検索要求に，世界知識を利用しないと展開されない単語が含まれているもの．
　例：「多角化事業の低迷」（多角化事業とは何か？）

BMIR-J1では，Ｆ１～Ｆ５のシステム機能を評価する検索要求をできるだけ均等に含まれるように６０件選択した．しかし，各検索要求に対する正解集合を求める過程で，正解を決めるためには，複数のシステム機能が必要となることが分かり，また，各システム機能の間には単純な包含関係は存在しないことが分かった．このため，BMIR-J1では，検索要求をＦ１～Ｆ５の５種類に分類せず，正解判定において，５種類のシステム機能のどれを必要としたかをマークすることにした．

検索要求は１文とする制限を設けたため，各検索要求には，検索の意図をできるだけ明確にするためのコメントを付記した．検索要求とコメントの例を以下に示す．この例は，検索要求ID１３（バージョン０）の検索要求を示しており，F=oooox は，正解判定において，上の５種類のシステム機能のいずれ

が必要かを判断したものである。その結果を，F＝?????という形でマークした．?????の５桁のo/xは，F１～F５の各々を必要とするか否かを表わす．ここで必要かどうかは，正解／不正解の判定に必要かどうかを意味し，次節で述べる正解のランク分けを行うための機能を含めない．N-1～N-3の３行は，検索要求に関するコメントである．

0013-0:Q:F=oooox:「千人以上の人員削減を行なう企業」
0013-0:Q:N-1:千人以上の人員の削減を予定している企業に関する記事を探す。
0013-0:Q:N-2:「行なう」の解釈として、今後実施予定のものを正解とする。
0013-0:Q:N-3:すでに行なわれたもの（実績）は、不正解とする。

２．３　正解集合

BMIR-J1では，正解判定は，次の２段階のレベル分けを行った．

・Ａランク　：検索要求を主題とする記事
・Ｂランク　：記事の主題は検索要求と異なるが，検索要求の内容を少しでも記述している記事

また，不正解と判断したが，正解か否かの判定が微妙なものは，Ｃランクとして参考のために残した．但し，Ｃランクは，不正解であり，網羅されている保証はない．さらに，正解判定の根拠をできるだけ明らかにするために，コメントを付記した．正解は，検索要求作成者が検索要求に付与したコメントを参考に，２名で判定したが，予想以上にペア間の正解判定のバラツキが大きかったため，検索要求作成者が，２名の正解判定結果，及びワーキンググループでの議論を参考に最終的に判断した．ばらつきが大きい検索要求は，人間でも正解判定に迷うものであり，検索システムにとっても正解を検索することは難しいと予想される．これらは，検索要求中の言葉の意味に多義性があるもの（「CD」を経済に関する記事では「キャッシュディスペンサー」，音楽に関する記事では「コンパクトディスク」と解釈），言葉のインスタンスへの展開の範囲が曖昧なもの（「ディスカウンター」，「トップ」，「コンピュータメーカー」），言葉の定義が不明確なもの（「外国企業」，「逆輸入」，「安売り」）などであり，正解判定に個人差がでる．この正解集合を求める過程で分かった検索要求の曖昧性は，できるだけ解消するように，検索要求のコメントに言葉の意味，展開範囲，定義などを追加した．

前節で例を示した検索要求ID１３の正解集合を下に示す．正解集合は，記事IDとそのランク，及びコメントからなる．

0013-0:R:09010003:A:「希望退職」＝人員削減
0013-0:R:09040130:A:東急百貨店による１０００人の削減計画
0013-0:R:09060035:C:新日鉄による２万人削減実績
・
0013-0:R:12080128:B:ＮＴＴによる３万人以上の削減計画、主題は子会社への業務移管
0013-0:R:12150175:A:本田技研工業による３０００人の削減計画

２・１で例を示した記事（ID=09010003）は，検索要求「千人以上の人員削減を行なう企業」に対して，Aランクと判定した．記事中には，キーワード「人員削減」は出現しないが，「希望退職募集」は,「人

員削減」と同等とみなせ，今後の実施予定を主題とした記事であるためAランクと判定した．Cランクの記事（ID=09060035）は，数値レンジは適合しているが，人員削減を行った実績を述べたものであり，検索要求の意図「人員削減の今後の実施予定」と異なるため，Cランク（不正解）と判定された。また，Bランクと判定した記事には，その記事の主題をコメントに記した．

BMIR-J1の検索要求とAランク，Bランクの正解文書数を付録に示す．検索要求ID＝４，９，１７，２１，２７，２８，２９，３１，３５，３６，４３，４７，４８，５９の１４件の検索要求では，正解文書数（AランクとBランクの合計）が，目標とした正解集合の最小文書数５件，最大文書数３０件の範囲を満たさないが，テスト版では，厳密な統計学的な理論からよりも作業量を優先してこれらの値を決めたため，これらの検索要求も残した．Aランクの正解文書数は延べ３３４件，Bランクは延べ２８１件であり，正解集合の文書数は延べ合計６１５件である．

3．検索要求の分類の試み

２・２節で述べたように，各検索要求の正解判定において，５種類のシステム機能のいずれを必要とするかを判断した．その結果は付録に示す通りである．BMIR-J1は，キーワードの論理演算に基づく検索システムから，統計処理，ニューラルネット，自然言語処理，知識処理など様々な検索手法の評価に利用されることを目標に作成した．各検索手法によって，正しく処理できるシステム機能は異なるため，検索要求は分類されている方が，本ベンチマークを利用し易いと考えられる．そこで，検索要求に付けたF=????? のパタンによって検索要求を以下のA～Fという６グループに分類することを提案する．

・グループA：基本機能のみ：	F=oxxxx：検索要求文ID=１～１０の１０件
・グループB：数値レンジ機能必要：	F=?o??? ：検索要求文ID=１１～１５の５件
・グループC：構文解析機能中心 ：	F=?xoxx：検索要求文ID=１６～２１の６件
・グループD：言語知識利用中心 ：	F=?x?ox：検索要求文ID=２２～３３の１２件
・グループE：世界知識利用中心 ：	F=?x?xo：検索要求文ID＝３４～４３の１０件
・グループF：言語知識と知識処理の併用：	F=?x?oo：検索要求文ID＝４４～６０の１７件

勿論，このグループ分けは，検索要求に対し正解集合を求めるために必要と考えらる機能で分類したものであり，各グループの機能を持たない検索システムの評価には本ベンチマークを利用できないという意味ではない．本グループ分けの利用法として，例えば，グループA，C，Dに分類された検索要求を対象に，自然言語処理を利用した検索手法と統計処理に基づく検索システムの適合率・再現率を比較することによって，両手法の優位性を判定することが可能である．情報検索システム相互の比較には，符号検定を用いることを提案している[8]．

4．おわりに

本稿では，日本語テキストを対象とした初めての情報検索システム評価用ベンチマーク BMIR-J1 について，その構成要素である対象文書，検索要求，正解集合について報告した．BMIR-J1の特徴 は，各検索要求に，正解集合を求めるために必要されるシステム機能（F１～F５）をマークしたことである．また，そのシステム機能のパタンによって，検索要求を分類することも試みた．これらの特徴によって，BMIR-J1 はテスト版のため対象文書こそ少ないが，様々な検索手法の優位性比較に広く利用されることを期待する．

今後は，BMIR-J1のモニタ公開により意見を集めて，日本語テキストを対象とした本格的な情報検索システム評価用ベンチマークの作成を進める予定である．

尚，BMIR-J1に関するお問い合わせ先は，以下の通りです．

bmirwg@ipsj.or.jp，NTTデータ通信（株）情報科学研究所，木谷（きたに）Tel: 044-548-4606

謝辞

本ベンチマークにおける新聞記事の利用を許可頂いた（株）日本経済新聞社に感謝いたします．また，BMIR-J1の作成を支援頂いた，データベースシステム研究会，田中克己主査に感謝いたします．

参考文献

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

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

[3] D. Harman, editor. Overview of The 3rd Text REtrieval Conference (TREC-3). National Institute of Standards and Technology, 1994.

[4] D. Harman, Overview of the 4th Text REtrieval Conference (TREC-4). the handout in TREC-4, National Institute of Standards and Technology, 1995.

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

[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] 木本晴夫ほか，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.

[8] 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.

[9] 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.

[10] 小川泰嗣ほか，日本語情報検索システムのためのベンチマークの構築，情報処理学会研究会報告，Vol.DBS100, pp.145-152,1994.

付録

検索要求ID	ファンクション分類 F1	F2	F3	F4	F5	検索要求	正解文書数 Aランク	Bランク
1	○	×	×	×	×	菓子メーカー	7	2
2	○	×	×	×	×	国内航空大手３社	9	4
3	○	×	×	×	×	任天堂	1	4
4	○	×	×	×	×	農薬	0	4
5	○	×	×	×	×	飲料品	4	10
6	○	×	×	×	×	液晶	2	5
7	○	×	×	×	×	ビデオデッキ	1	4
8	○	×	×	×	×	携帯電話またはパーソナルハンディホン	2	6
9	○	×	×	×	×	会社更正法	2	2
10	○	×	×	×	×	減税	3	9
11	○	○	×	×	○	開始時間が午前１０時の日経ビジネススクール	13	0
12	○	○	○	○	×	３期以上連続の減益企業	3	5
13	○	○	○	○	×	千人以上の人員削減を行なう企業	17	5
14	○	○	○	×	×	中国にある資本金五億円以上の合弁企業	10	0
15	○	○	○	×	○	１ドル＝１０５円以上の円高	5	5
16	○	×	○	×	×	半導体製品の生産	1	7
17	○	×	○	×	×	電話通話料金の値下げ	1	2
18	○	×	○	×	×	所得税の減税	1	6
19	○	×	○	×	×	コンピュータメーカーの人員削減	1	5
20	○	×	○	×	×	非製造業による現地法人設立	5	0
21	○	×	○	×	×	電力業界の自民党に対する献金	1	0
22	○	×	○	○	×	製販一体化	8	5
23	○	×	○	○	×	円高による物価の低下	1	8
24	○	×	○	○	×	冷夏の被害	3	11
25	○	×	○	○	×	メーカーの減益に対する対策	17	3
26	○	×	○	○	×	株価動向	3	4
27	○	×	○	○	×	ファクシミリの市場動向	1	0
28	○	×	○	○	×	日本製品の対米輸出量の実績	0	3
29	○	×	○	○	×	企業における情報共有の導入事例	1	1
30	○	×	○	○	×	業績悪化を原因とする企業の合併の事例	10	4
31	○	×	○	○	×	日本の製造業における生産性向上またはコスト・ダウンの事例	6	25
32	○	×	○	○	×	銀行の経営計画	13	2
33	○	×	○	○	×	リエンジニアリングかリストラの定義	4	4
34	○	×	○	×	○	多角化事業の低迷	4	7
35	○	×	○	×	○	異業種会社間の共同経営	1	1
36	○	×	○	×	○	特殊装備自動車の新会社設立	2	1
37	○	×	○	×	○	電気メーカーの中国への投資	13	0
38	○	×	○	×	○	外国企業の日本への進出	6	1
39	○	×	○	×	○	権限の役員への委譲	3	2
40	○	×	○	×	○	管理部門の統廃合と営業部門の強化を行なう会社	4	2
41	○	×	○	×	○	アジア諸国による物資または製品の日本への輸出	6	17
42	○	×	○	×	○	北陸地方の会社	4	5
43	○	×	○	×	○	第三セクター事業運営	1	1
44	○	×	○	○	○	現地調達	14	3
45	○	×	○	○	○	流通改革	6	8
46	○	×	○	○	○	経営陣刷新	7	2
47	○	×	○	○	○	女性の雇用問題	1	3
48	○	×	○	○	○	企業の社会貢献	2	2
49	○	×	○	○	○	企業の低価格競争	11	16
50	○	×	○	○	○	第３次産業のサービス向上	12	4
51	○	×	○	○	○	逆輸入を行なう日本企業	7	0
52	○	×	○	○	○	安売りを行なう流通業者	12	5
53	○	×	○	○	○	トップの不況対策に関する発言	12	5
54	○	×	○	○	○	業績不振の責任を取った経営者	1	6
55	○	×	○	○	○	企業による配下企業の再編成	9	5
56	○	×	○	○	○	円高対策のためのメーカーの海外進出	13	8
57	○	×	○	○	○	行政機関が関係する不況対策	10	3
58	○	×	○	○	○	不況におけるディスカウンターの台頭	4	3
59	○	×	○	○	○	買い取り制による低価格ブランドの成長	1	2
60	○	×	○	○	○	経営多角化の事例	12	13