論　文

不完全な評価セットに対する検索エンジンの性能評価指標の開発

大平　倫宏[*1)]　　富山　真一[*2)]

Development of metrics for quality evaluation of information retrieval systems for incomplete evaluation data set

Norihiro Ohira[*1)], Shinichi Tomiyama[*2)]

Until now, numerous information retrieval (IR) systems have been developed and utilized. Nevertheless, it is uncommon that they are objectively compared with other systems. In academic studies, the qualities of IR systems are numerically measured based on a complete evaluation set that gives relations between search results and query keywords. However, these methods have never been popular in general, because they cost too much to evaluate the relevance of all the relations by hand.

In this study, we developed two methods to evaluate IR systems using an incomplete evaluation set. The first method uses machine learning. We have obtained favorable results in experiments using only 50 manually-evaluated relations. The second is a new metric for the quality of IR systems. It differs from conventional inf-nDCG in the point that it can handle the query intent. In experiments, it was more robust against the incomplete evaluation of relationships than inf-nDCG.

キーワード：検索エンジン，情報検索，ビッグデータ，評価指標，ランキング

Keywords：Search engine, Information retrieval, Big data, Metric, Ranking

1.　はじめに

高度情報化社会の進展に伴い，数多くの情報検索方法が提案されている。中小企業に関しても，ベンチャー企業などで情報検索システムの開発・提供を主要な業務とする企業が増えつつある。しかし，市場に供されている検索システムが，他のシステムなどと比較して，優れている点を客観的に主張していることは稀である。その性能の評価は，開発・提供者，利用者らの主体的な判断に委ねられている場合が多い。

一方，情報検索システムの性能を客観的に評価するために学術研究で主に利用されている方法として，検索結果と検索キーワードとの関連性をあらかじめ人手を用いて定める方法がある。しかし，人手を用いて多くの関連性を定めることはコストがかかり，中小企業等には負担が大きい。このことが，検索システムの客観的な評価が行われない大きな理由となっている。

検索結果ランキングの代表的な評価指標として，nDCG[(1)]があり，広く利用されている。それに対して，近年は検索システムを利用する者の検索意図（query intent）をランキングに反映する指標が着目されている。例えば，検索キーワード「書籍」で検索した場合に，その時の人気の書籍1冊だけが上位にランキングされた検索結果を返すシステムがあった場合，これはnDCGでは高く評価される。しかし，実際の利用者は，同じ書籍の情報ばかりを得たいと思うことは稀であり，nDCGでの評価が必ずしもそぐわない。この問題を踏まえ，Clarkeらは，α-nDCGを提案した[(2)]。しかし，評価セットとして多段階評価したものを利用できないことや，実用上では正規化（ここでは，評価値の値を0から1の範囲に収めることとする）できないことが問題として挙げられている[(3)]。さらに，SakaiらによってD-measures[(3)]が提案されており，検索意図の考慮が検索結果評価指標の重要な特性の一つとなっている。

一方，実際には，人手を用いるにはコストがかかるため，全ての検索結果に対してではなく，一定数の検索結果と検索キーワードとの関連性だけについてあらかじめ人手を用いて決めておき（以下，このように決めた関連性評価を不完全な評価セットとする。逆に完全な評価セットは，全ての検索対象に対して，評価を定めているものとする），それに基づき検索結果ランキングを評価する方法も提案されている。Buckleyらはbprefを提案したが[(4)]，ロバスト性がないという評価を受けている[(5)(6)]。Sakaiらは，bprefに代わる評価指標として，関連性の定められていない検索対象を除外することによって圧縮したランキング結果を利用するCondensed-list metricsを提唱した[(5)(6)]。また，Aslamらは無作為抽出法によって，評価値を推定する方法を提唱したが[(7)]，その複雑さのため広く利用されてはいない[(8)]。それを踏まえEmine，Aslamらは，より簡単な手順によって評価値を求めることが可能な指標であるinfAPを提案した[(8)]。さらに，実

事業名　平成26年度　基盤研究「完全な評価セットに対する検索エンジンの性能評価指標の開発」

[*1)] 生活技術開発セクター
[*2)] 情報技術グループ

際の利用における状況を考慮して無作為抽出だけによらないように改良した評価指標であるxinfAPと，それと同様の特徴を備えたnDCGの拡張版であるinfNDCGを提案した[9]。

上述のように，検索エンジンに対する評価指標は多く提案されているが，検索意図を考慮し，かつ，不完全な評価セットに対応した指標は存在しない。本研究の目的の一つは，そのような特徴を備える評価指標を新たに開発することである。

また，機械学習などの統計的手法を用いて，関連性が既知の既存の評価セットから，評価が未知である検索対象に対する関連性評価を推定する方法についても開発を行った。こちらの方法については，検索意図の考慮は行わないが，既知の評価データセットが小さい場合でも活用可能であることを目的としている。

これらの研究を行うことで，中小企業などでも大きな負担をかけずに，検索システムの客観的性能評価ができるようになることを目指した。

本稿では，2.1節で機械学習を用いた未知文書の評価方法について説明する。また，2.2節では，検索意図を考慮し，かつ，不完全な評価セットに対応した評価指標を新たに開発したので，それについて説明する。

2. 研究内容

2. 1 機械学習を用いた方法　以下のような手順で機械学習によって未評価の文書の関連性評価を推定した。本研究では，基本的に文書を対象として，検索ランキングを作成すると仮定している。文書以外を検索対象とする場合は関連性評価の方法を変更する必要があるが，同様に適用可能である。

(i) 全文書に対して，各文書に現れる各単語の数を集計する。
(ii) 各文書の各単語に対するTF-IDF特徴量（単語の出現頻度・単語の重要度）を計算する。
(iii) 評価済み文書のTF-IDF特徴量を訓練データとして，教師付き機械学習を行う。
(iv) 未評価の文書について，学習結果を元に関連性の判定を行う。

ここでTF-IDF[11]は，

$$\mathrm{TF-IDF}(t,d,D)=\mathrm{TF}(t,d)\cdot\mathrm{IDF}(t,D) \quad \cdots\cdots (1)$$

$$\mathrm{TF}(t,d)=\frac{n_{t,d}}{\sum_k n_{k,j}} \quad \cdots\cdots (2)$$

$$\mathrm{IDF}(t,d)=\frac{\log|D|}{\left|\{d\in D;t\in d\}\right|} \quad \cdots\cdots (3)$$

の式によって計算する。Dは全文書の集合，$|D|$は総文書数，$n_{t,d}$は単語tの文書dにおける出現数，$\left|\{d\in D;t\in d\}\right|$は単語$t$を含む文書の総数を表す。TF（Term Frequency）は単語の出現頻度，IDF（Inverse Document Frequency）は単語の特殊性を表している。

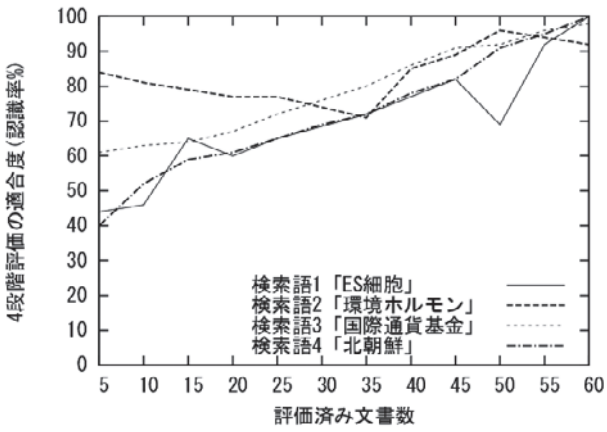

図1. 機械学習に利用する評価済み文書数と4段階の関連性評価の適合度

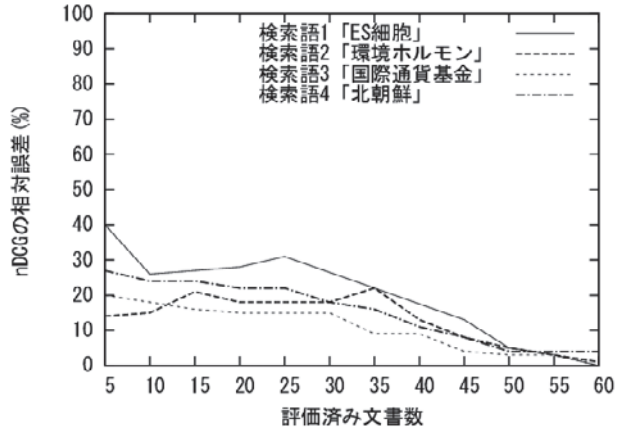

図2. 機械学習に利用する評価済み文書数とnDCGの相対誤差の関係

本手法の有効性を確認するために以下のような状況で実験を行った。利用するデータとして，NTCIR（NII Testbeds and Community for Information access Research）[11]のNTCIR-6 CLIRにおける評価データセットを用いた。評価データセットでは，読売新聞データベース2000年99,207文書，毎日新聞データベース2000年306,709文書について，検索語に対しての関連性を「関連性が高い」,「関連性がある」,「関連性が低い」,「関連性が無い」の4段階で評価している。全文書中の重複を除いた単語の総数は228,297語であり，そのそれぞれについてTF-IDF特徴量を算出した。機械学習には，オープンソースの機械学習ライブラリであるLIBSVM[12]のサポートベクターマシンを利用した。各文書の特徴量としては，各単語のTF-IDF特徴量を並べた228,297次元のベクトルを用いた。

開発した手法において，機械学習に利用する関連性評価済み文書数と推定される評価の適合率の関係を表した図が，図1である。それぞれの検索語に対して，K=10として，K-分割交差検定を行い10回の試行の平均を算出している。また，適合率の算定では，4段階で推定した関連性評価値が，あらかじめ定められた真の関連性評価値と一致したときを適合と判定している。図1では基本的に機械学習に利用する

文書数が増加するにつれ，推定値の適合度が向上している。

さらに，得られた推定結果から，従来から使用されている検索ランキングの評価指標であるnDCG[1]を用いて，完全な評価セットに対する評価値（真値）と不完全な評価セットを用いた評価値（推定値）との相対誤差を求めたのが，図2である。ここでnDCGは，

$$\mathrm{nDCG} = \frac{\mathrm{DCG}}{\mathrm{IDCG}} \quad \cdots\cdots (4)$$

$$\mathrm{DCG} = \sum_{r=1}^{z} \frac{2^{rel_r}}{\log_2(r+1)} \quad \cdots\cdots (5)$$

の式によって計算した。rはランキング順位を表しており，rel_rはランキングrの文書の関連性を，ZはnDCGの評価に利用する順位の最大値を表している。また，IDCGは，理想的なランキング（全対象を関連性の高い順に並べた場合）に対してのnDCGの値である。

図1，2では，基本的に評価済み文書数が多いほど相対誤差が小さくなっており，利用する評価済み文書数が50文書以上であれば，真値との相対誤差が5%以下となっている。50対象の評価程度であれば，中小企業などのコスト余力のないグループにとっても十分に用意可能である。また，図1の4段階評価の適合度に比べると，図2のnDCGの相対誤差はより良いように見える。このことは，図1では，関連性評価が1つでも違えば間違っていると判定しているため，厳しく評価されているが，実際には推定値は真値と近いため，実際のnDCGでは精度が良くなる結果となっていると考えられる。

2. 2　不完全な評価セットに対する評価指標の開発

inf-nDCG[9]では，ランキング対象全体を，共通部分を持たない複数の層に分けることを考慮する。層ごとに標本（検索対象）の抽出確率を変化させて，検索対象のグループを作成する。そのグループに対してランキングを作成して，評価することを考える。この考えは，例えば，TREC（Text REtrieval Conference）[13]のTerabyte 2006などでの効果的な利用を想定している。Terabyte 2006においては，対象とする文書数が非常に多いため，評価セットを3層に分けており，ある層では多くの文章に対して高い関連性が定められており，別の層では関連性の評価付けが不完全であるという状況にある。これに対して，inf-nDCGでは，例えば，高い関連性を持つ文書を多く持つ層からの抽出確率を高めることで，関連性の高い文書を多く抽出し，確率的な方法によりながらも，高いロバスト性を実現することが可能である。

それを踏まえ，従来の指標であるD-measures[3]を基に，inf-Dmesuresと呼ぶべき指標を考えることとする。D-measuresは，システムの利用者は，様々な検索意図に基づいて検索すると仮定しており，一つの検索語に対して複数の検索意図が該当する状況に対応している。例えば,「水」を検索語とすると,「飲料水」や「雨水」,「ミネラルウォーター」などの検索意図が考えられる。実際の指標値の計算は，

$$\mathrm{D-measure} = \sum_{r=1}^{z} \frac{GG(r)}{\log_2(r+1)} \quad \cdots\cdots (6)$$

$$GG(r) = \sum_i \Pr(i \mid q) g_i(r) \quad \cdots\cdots (7)$$

によって行う。ここで，$\Pr(i \mid q)$は，検索語qに対する検索意図iの該当確率，$g_i(r)$は順位がrである文書の検索意図iとの関連性とする。

本研究では，inf-nDmesuresとして，以下のような指標値を提案する。まず，各文書dについて，

$$GGD(d) = \sum_i \Pr(i \mid q) g_i(d) \quad \cdots\cdots (8)$$

のように定める。

次に，inf-IDmesuresを，

$$RD(\zeta) = \frac{r_s(\xi \mid \xi \le GGD(d) < \xi + \Delta)}{n_s} N_s \quad \cdots\cdots (9)$$

$$\hat{RD}(\xi) = \sum_{\forall s} \hat{RD}_s(\xi) \quad \cdots\cdots (10)$$

の式を利用して計算する。各層sに対して，$r_s(\xi \mid \xi \le GGD(d) < \xi + \Delta)$は抽出した文書の中で$\xi \le GGD(d) < \xi + \Delta$の条件を満たす文書の数を表しており，$\Delta$は事前に定める評価値の分割に利用する値である。$N_s$は各層で抽出した総文書数を，$N_s$は各層の総文書数を表している。式(9),(10)から，文書全体での$\xi \le GGD(d) < \xi + \Delta$の条件を満たす文書の数の推定値は$\hat{RD}(\xi)$で求められる。これらを用いて，理想的なランキングにおけるD-measuresの推定値を求め，その値をinf-IDmesuresとする。

次に，inf- Dmesuresについては，

$$\text{inf-nDmesures} = \sum_{\forall s} \frac{Z_s}{Z} \cdot E\left[x_r \mid \text{document at rank } r \in s\right] \quad \cdots\cdots (11)$$

$$E\left[x_r \mid \text{document at rank } r \in s\right] = \frac{1}{n_s} \sum_{\forall i \in \text{sampled}_s} Z \cdot \frac{GG(r)}{\log_2(r+1)} \quad \cdots\cdots (12)$$

の式によって求める。Z_sは検索システムがランキング出力した文書で層sに含まれるものの総数を表し，Z_s/Zの確率で層sから文書を無作為抽出するとする。また，sampled$_s$は層sから抽出された文書の集合とする。

最後に，

$$\text{inf-nDmesures} = \frac{\text{inf-Dmeasures}}{\text{inf-IDmeasures}} \quad \cdots\cdots (13)$$

を用いて，inf-nDmesuresを計算する。

inf-nDmesuresと他の指標値との比較を行った。評価にはNTCIR-9 INTENT（INTENT-1）日本語ドキュメントランキングテストコレクション[11]における関連性評価データを利用した。そこでは，6,700万件の日本語ウェブページに対して，100組の検索語とそれらの検索意図の該当確率$\Pr(i \mid q)$と各文書の各検索意図に対する関連性を5段階で評価している。

文書と検索語自体との関連性は定められておらず，inf-nDCGとrandom sampling nDCGはそのままでは計算できないため，文書と検索語自体との関連性は，検索意図の関

表1.　従来の指標との比較

	nDCG	inf-NDCG	D-mesures	inf-nDmeasures
検索意図への対応	×	×	○	○
計算量	○	○	○	○
不完全な評価セットへの対応	×	○	×	○
不完全な評価セットでのロバスト性	×	△	×	○

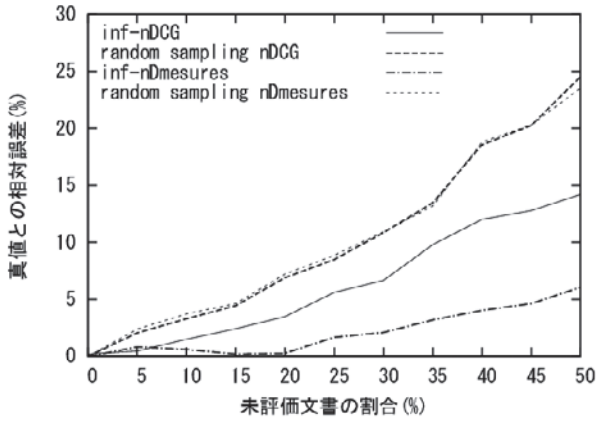

図3.　指標値ごとの未評価文書数の増加に対する
真値からの乖離の様子

連性の中で最大のものを用いた。random sampling nDCGとrandom sampling nDmesuresは無作為抽出を利用して，評価されていない文書は関連性がないとして計算した。

inf-nDCGとinf-nDmesuresについては，層を2つに分け，一つ目の層には関連性評価データから関連性の高い文書を順におき，二つ目の層には評価されていない文書を含むそれ以外の文書を配置し，一つ目の層の選択確率を90%とした。また，inf-nDmesuresについては，評価値が100段階評価となるように式（14）のΔを選択した。このようにして，未評価文書がない場合にランキングが理想的に並べられていると仮定して，その真値との相対誤差を，100組の検索語に対して平均をとったものが，図3である。

提案する指標値inf-nDmesuresでは，図3から，従来の指標値inf-nDCGよりも相対誤差が低いことがわかる。これは，inf-nDmesuresでは関連性評価の値を100段階としているため，inf-nDCGの5段階評価より粒度が小さく，より正確に推定可能であるためと考えらえる。また，表1は提案指標と従来の指標との比較を表しており，従来指標と比べて優れていることがわかる。

3.　まとめ

本研究では，検索システムの性能評価に関して，大きな負担を行うことが難しい中小企業等の組織においても，有効である性能評価方法を提案した。これらの方法によって，検索システム開発全体の活性化が行われることを期待する。

また，今回の実験では，実際の検索エンジンを使用しての比較は行わなかった。実用上の問題を調査する意味でも，そのような比較を行うことが今後の課題である。

（平成27年7月13日受付，平成27年7月28日再受付）

文　　献

(1) Kalervo Jarvelin and Jaana Kekalainen, “Cumulated Gain-Based Evaluation of IR Techniques”, ACM Transactions on Information Systems, Vol.20, No.4, p.422-446 (2002)
(2) Charles L.A. Clarke, Maheedhar Kolla, Gordon V. Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher and Ian MacKinnon, “Novelty and Diversity in Information Retrieval Evaluation”, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, pp.659-666, 20-24, July, 2008
(3) Tetsuya Sakai and Ruihua Song, “Evaluating Diversified Search Results Using Per-Intent Graded Relevance”, Proceedings of the 34th international ACM SIGIR Conference on Research and Development in Information Retrieval, Beijing, China, pp.1043-1052, 24-28, July, 2011
(4) Chris Buckley and Ellen M. Voorhees, “Retrieval Evaluation with Incomplete Information”, Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Sheffield, United Kingdom, pp.25-32, 25-29, July, 2004
(5) Tetsuya Sakai, “Alternatives to Bpref”, Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Amsterdam, The Netherlands, pp.71-78, 23-27, July, 2007
(6) Tetsuya Sakai and Noriko Kando, “On Information Retrieval Metrics Designed for Evaluation with Incomplete Relevance Assessments”, Information Retrieval, Vol.11, No.5, pp.447-470 (2008)
(7) Javed A. Aslam, Virgil Pavlu and Emine Yilmaz, “A Statistical Method for System Evaluation Using Incomplete Judgments”, Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, USA, pp.541-548, 06-11, August, 2006
(8) Emine Yilmaz and Javed A. Aslam, “Estimating Average Precision with Incomplete and Imperfect Judgments”, Proceedings of the 15th ACM International Conference on Information and Knowledge Management, Arlington, Virginia, USA, pp.102-111, 06-11, November, 2006
(9) Emine Yilmaz , Evangelos Kanoulas and Javed A. Aslam, “A Simple and Efficient Sampling Method for Estimating AP and NDCG”, Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Singapore, Singapore, pp.603-610, 20-24, July, 2008
(10) Anand Rajaraman and Jeffrey David Ullman, “Mining of Massive Datasets”, Cambridge University Press, New York (2011)
(11) NTCIR, http://research.nii.ac.jp/ntcir/index-ja.html
(12) LIVSVM, http://www.csie.ntu.edu.tw/~cjlin/libsvm
(13) TREC, http://trec.nist.gov/