解説

Web 検索における検索結果ランキング ～評価手法とアルゴリズム～†

清水 徹*

1. はじめに

1960年代にJ. C. R. Lickliderらが構想したコンピュータネットワーク技術[Licklider68]は1990年代にインターネットおよびWorld Wide Webとして爆発的に普及し、人々の生活に大きな影響を及ぼすようになった。また同時に、ネットワークを通じてアクセス可能な情報の量も指数関数的に増加を続けている。今日、膨大なWeb上のドキュメントから求めるものを探し出すためには、情報学の分野で長く培われた全文検索技術を実装しWebに適用した「Web検索エンジン」が必要不可欠となっている。

日々、多くのインターネットユーザが利用する商用Web検索サービスだが、検索エンジン本体の開発・運用を行っているプレイヤーの少なさもあってか、その開発や性能評価の有り様について公に流通している知見は現状それほど多くはない。筆者は国内の主要な検索サービスプロバイダーの一つ、Yahoo! JAPANにおいてWeb検索エンジン開発に携わっており、本稿は商用Web検索エンジンの性能評価について、中でも最も重要とされる「関連性」に重点を置きながら、それを支えるランキング・スコアリングの一般論と併せて解説を試みるものである。

なお、ドキュメントのクロールや転置インデックスの作成、ポスティングリストのマージソートといった全文検索技術の根幹をなすその他の要素は、本稿では扱わず、これらについては本分野の標準的なテキスト[Manning08, Croft09]を参照されたい。

2. Web 検索エンジンの性能評価フレームワーク

現在広く使われているWeb検索サービスとしては、いわゆる「検索窓」にクエリを入力すると、それに該当するWeb上のドキュメントについての「タイトル」「サ

マリ」「URL」の組からなる検索結果のリストが返される、という形が一般的である。さて、このようなサービスにおける「良いエンジン」「良い検索結果」とはいったいどのようなものだろうか。

2.1 非機能面の評価

システム開発の要件定義工程においては「機能要件」「非機能要件」がそれぞれ定められ、それらを満たすべく開発が進められる。オンラインシステムの一つとして検索エンジンを捉えたとき、非機能の側面については当然ながら、他の大規模なWebアプリケーションとの間で大きな違いはない。すなわち

- ・ Availability (可用性)
 - － システムが停止することなく、安定して稼働を続けているか
- ・ Latency (応答時間)
 - － クエリ投入から検索結果表示までの所要時間が十分に短い
- ・ Throughput (単位時間あたりの処理量)
 - － 一定時間に処理可能なクエリ数が十分に大きい

といった観点で高い性能を保つことが要求される。

Web検索サービスの運営においては「情報流通インフラとしての公共性」の観点などから、実質的に100%の可用性が要求され、Web検索サービスを提供する各社においてはその実現のため真摯な努力が続けられている。

2.2 機能面の評価

筆者の携わる Yahoo! JAPANのWeb検索サービスにおいては、以下の4つを主要な観点として機能評価を行っている。

- ・ Relevance (関連性)
 - － クエリの意図と合致した検索結果が提供されているか

† Evaluation Methods and Algorithms for Web Search Ranking
Toru SHIMIZU

* ヤフー株式会社 R&D統括本部 プラットフォーム開発本部
Platform Development Department, R&D Division, Yahoo! JAPAN Corporation

- クエリの意図に、より良く合致した検索結果がより上位に表示されているか
- ・ Comprehensiveness (網羅性)
 - 検索対象の母集団として十分な量のドキュメントがインデキシングされているか
 - 個々のクエリに対し十分な量の検索結果が提供されているか
- ・ Freshness (新鮮度)
 - 検索対象となるドキュメントが十分な鮮度を保っているか
 - Web上に新規に現れた重要なドキュメントが、速やかに検索結果に供されるか
- ・ Presentation (表示の適切さ)
 - UI全体として見やすい、使いやすいものとなっているか
 - クエリの意図に合致した検索結果を見つけやすいか

comprehensivenessを表す指標としては、インデックスにおける全ドキュメント数、一定数以下の検索結果数しか持たないクエリの割合などがあり得る。freshnessについては、検索結果上に現れるドキュメントのクロール日からの平均日数や、インデキシング日からの平均日数などを用いることが考えられる。

クロールおよびインデキシングに使えるコンピューティングリソースが一定であるとする、一般的にはcomprehensivenessとfreshnessの間で直接のトレードオフが発生することになる。大規模なドキュメントデータベースは更新に日数を要してしまう一方、Webスケールのドキュメントデータベースを小規模に留めることはそもそも望むべくもなく、この相反する二軸について何らかの折り合いをつけなければならない。しかしながら、例えば、ドキュメント毎に検索結果に表れる回数をカウントすると、カウント数とそれに対応するドキュメント数の間にべき乗則が成り立つことがよく知られている。これは多数回検索されるような重要ドキュメントは全体の中のごく少数であり、重要度に応じた更新リソースの傾斜配分が有効であることを示唆している。

また、非機能面の評価指標として述べたlatencyと、freshnessの間のトレードオフについても注意が必要となる。全文検索エンジンも広義にはデータベースシステムであるため、RDBMSなどと同じくクエリとレスポンスの組についてのキャッシングが有効であり、キャッシュヒットの有無によりlatencyが大きく異なる。キャッシュヒット率を上げることでlatencyは改善するが、その結果、新しく検索データベースに

追加されたドキュメントがユーザに提供されるタイミングが全体的に遅れ、freshnessは低下することになる。

Web検索のユースケースである「数十億、数百億のWebドキュメントの中から、クエリの意図に合致した少数の検索結果を選び出す」を適切に実現するために、relevanceは常に重要視され、測定・改善のために大きなリソースが充てられている。

3. Relevance 評価手法の概要

本節では、Web検索エンジンのrelevance評価における代表的な手法として、クエリ・ドキュメントペア単位のrelevanceについて多数の判定結果を収集し、定量指標に落とし込む一連の流れを紹介する。

3.1 ガイドライン

relevance測定のためには、まず予めどのようなクエリに対してどのような検索結果が望ましいか「ガイドライン」を策定しておくことが必要である。ガイドラインとは、クエリの意図するところと完全に合致するとみなせるドキュメントとはどのようなものか、ほぼ合致するものとは、ある程度合致するものとは、わずかに合致するものとは、といった合致度合いの段階的な基準を定め、明文化したものである。

3.2 クエリ・ドキュメントペアに対する評価

このひとつひとつの合致度合い(relevance)の段階をここではgradeと呼ぶこととし、5～0の6段階のgradeがあると考えよう。あるクエリと検索結果ドキュメントのペアが与えられたとき、それを評価して、ドキュメントが完全に検索要求に合致すれば(ガイドラインに定められたその基準を満たせば)gradeは5となる。ほぼ合致するものであればgrade 4となり、grade 1の基準にも満たず全く検索タスクに貢献しないドキュメントのgradeは0となる。

さらに、評価結果をより定量的な重要度スコアとして扱えるようにするため、重要度スコア値をscoreと名づけた上で、gradeからscoreへのマッピングを定義しておこう。ここでは例としてgrade r のクエリ・ドキュメントペアのscore値 s が

$$s = 2^r - 1 \quad (1)$$

で与えられるものとする。最も関連性の高いgrade 5に対するscoreは31、grade 0であればscore 0となる。

具体例として「ピアノ 中古」というクエリを考えると、「中古ピアノ販売・買い取り専門サイトのトップページ」はおそらくgrade 5(score 31)を獲得するであろう。「中古ピアノのオークションページ」であれば

grade 4 (score 15) の評価が妥当であり、あるいは「とあるレストランの食事についてのブログエントリで、中古車で乗り付けたこと、店内のピアノのBGMが素敵だったことが言及されたため、たまたま検索結果にヒットしたもの」であれば基本的にクエリの意図と無関係とみなせ、grade 0 (score 0) に該当する。

このようなクエリ・ドキュメントペア単位のgrade付けは、ガイドライン整備や評価者のトレーニング、評価実施などで通常、大きな手間と時間を要するが、relevanceの定量的な取り扱いを可能にするためには必要不可欠であると言える。

3.3 クエリ単位の検索結果の評価

次に、あるクエリの検索結果について、第1位のドキュメントから一定順位(depth n)までgradeを付与してみよう。このscoreの合計値を単純にクエリ単位の評価値とする指標がCG (Cumulative Gain) である。順位 k のドキュメントのgradeを $r(k)$ とすると以下のよう

$$\sum_{k=1}^n (2^{r(k)} - 1) \quad (2)$$

これでrelevanceの基準の一つ「クエリの意図と合致した検索結果が提供されているか」については計測できるが、もう一つの基準である「クエリの意図に、より良く合致した検索結果がより上位に表示されているか」について詳細な計測を行うことが出来ない。DCG (Discounted Cumulative Gain) はこれを考慮し、順位のlogのオーダーで評価への寄与が下がるよう係数を乗じた

$$\sum_{k=1}^n \frac{2^{r(k)} - 1}{\log_2(1+k)} \quad (3)$$

として定義される。depthの値を末尾に付け、例えば3位までを対象にした場合はDCG3、10位までであればDCG10と呼ぶ。

3.4 DCG評価・計算の具体例

例として、「バンクーバーオリンピック」というクエリの検索結果についてDCG3を用いた評価を行ってみる。手順としては、上位3位までの検索結果を評価しそれぞれにgradeを付与、それをもとにscoreを計算し、順位を考慮するための係数を掛けた上で足し上げを行えばよい。

同検索エンジンの異なるセッティング間や、あるいは異なる検索エンジン間で比較を行っているとして、図1のように2種類の結果が得られたとする。それぞ

バンクーバーオリンピック	検索
1. 熊谷刺繍センタートップページ	→ $r_1 = 0$
2. 金沢スポーツ新聞: モーグル特集	→ $r_2 = 3$
3. バンクーバーオリンピック公式サイト	→ $r_3 = 5$
→ DCG3のスコア: 19.9	

バンクーバーオリンピック	検索
1. バンクーバーオリンピック公式サイト	→ $r_1 = 5$
2. 金沢スポーツ新聞: モーグル特集	→ $r_2 = 3$
3. 熊谷刺繍センタートップページ	→ $r_3 = 0$
→ DCG3のスコア: 35.4	

図1 2つの検索結果についてのDCG3計算例

れ上位3位を構成する検索結果のドキュメントは変わらず、その順位のみが異なるような状況である。

前者では、「熊谷刺繍センター」という、検索要求との関連が全く明らかでないドキュメントが堂々1位を占めており、さらにパーフェクトな評価を得ている公式サイトが3位ということで、ランキングについての改善の余地が大きい。

後者は、この3つのドキュメントからなる検索結果ページとしては最良の順序となっており、直感を裏付けるような形でスコア的にも前者を大きく上回っている。加えて、「熊谷刺繍センター」の代わりに、冬季オリンピックと関連していて1以上のgradeが付くような他の何らかのドキュメントが来れば、スコアはさらに上昇することになる。

3.5 検索エンジン全体の評価

Web検索サービスに実際に投入されるクエリの分布を反映するような多数のサンプルクエリについて、先述したようなクエリ単位の評価を行い、DCG10等の評価値の平均を求めることによって該当検索エンジンのrelevanceの評価値が求められる。また当然、同一のガイドライン、同一のクエリセットを複数検索エンジンに適用することで検索エンジン間の比較評価が可能となる。

予めクエリに対してジャンルの区分を付与しておけば、エンターテイメントなど特定分野のクエリ群についてrelevanceを比較する、といったことも再集計により容易に実現できる。

4. ランキングアルゴリズム概要

前節におけるrelevance評価の手法を踏まえ、本節ではさらにWeb検索エンジンにおけるランキングとスコアリングについて概説する。

4.1 ランキングとスコアリング

Web検索エンジンにおけるランキングとは、検索実行時の処理において、クエリのキーワードにヒットした個々のドキュメントに対してrelevanceのスコアが付与され、それによってソートが行われた結果とみなすことができる。

情報検索におけるrelevanceのスコアリングの手法としては、索引語のドキュメント中の出現頻度およびドキュメントセット中の文書頻度をもとにしたTF-IDF [Jones72]のようなアプローチや、リンク構造からドキュメントの重要度を推定するHITS [Kleinberg98]、PageRank [Brin98]をはじめとして、さまざまなものが知られている。その他、個々のドキュメントがアダルト系のものであるか、スパムであるか、といった判定結果が数値として残されていれば、これらも同様にスコア的一种とみなせるであろう。

個々のドキュメントについて、あるいはクエリ・ドキュメントのペアについて多種多様なスコア値が付与されうる中で、これらの情報を包括的に扱い、relevanceの総合スコアを導き出すためにはどのようなアプローチを取ればよいだろうか。

ここで、非常に多くのクエリ・ドキュメントペアについてのrelevanceのスコア s が、3.2で述べたような評価によって既に得られていると仮定し、さらに、各クエリ・ドキュメントペアの(クエリ単独、ドキュメント単独で定まるものも含む)素性ベクトル x も得られていると仮定する。そうすると、それらの多数の素性とスコアの組

$$\{(\vec{x}_i, s_i)\}_{i=1}^n \quad (4)$$

をトレーニングデータとして、人による判断をできるだけ正しく再現するような関数 $s = F(x)$ を予めオフラインで機械学習的に求めておき、検索実行時のスコアリングに利用するようなアプローチが考えられる。関数 F のフィッティングについては、損失関数 $L(s, F)$ を導入し、その期待値を最小にするような F を求める回帰問題

$$F^* = \arg \min_F E_{s, \vec{x}} L(s, F(\vec{x})) \quad (5)$$

として定式化される。

このようなフレームワークにより、クエリの長さ、

図2 ランキングに機械学習を利用する検索エンジンアーキテクチャの例[Wikipedia10]

ドキュメントのサイズといった属性や、PageRank的なスコアおよびTF-IDFのようなワードスコアなど、多くの素性を素性ベクトルの要素として同列に扱いながら、総合的なrelevanceスコアを導くことが可能となる。

4.2 回帰問題を解くために

4.1で述べた回帰問題の具体的な解を与える手法として様々な選択肢がある中で、本節では、非常にシンプルかつノンパラメトリックな関数にも適用可能な解法であるGradient Boosting [Friedman01]を取り上げる。

ある微分可能な関数が与えられたとき、それが最小値を取るポイントを探したいが解析的に解けない場合は、勾配降下法を用いればよい。ある適当な箇所から探索をスタートし、各変数の偏微分から得られる「勾配最大」の方向の逆方向へ小さなステップを繰り返しながら、収束する箇所を目指すのである。

損失関数は関数の関数、すなわち汎関数であるが、偏微分の代わりに汎関数微分を用いることにより、同様のアプローチで期待値最小の解を探索できる。ある関数 f_0 からスタートし、損失関数の期待値を小さくするような差分の関数 f_k を足し上げていくことで F^* を求めるとすると、 F^* は

$$F^*(\vec{x}) = \sum_{k=0}^N f_k(\vec{x}) \quad (6)$$

である。個々の差分 f_k は、「勾配最大となる方向」に相当する関数 g_k 、「ステップの幅」を調整するための係数 ρ_k を導入して

$$f_k(\vec{x}) = -\rho_k g_k(\vec{x}) \quad (7)$$

と書ける。 g_k は具体的には、

$$g_k(\vec{x}) = \left[\frac{\partial E_s[L(s, F(\vec{x})) | \vec{x}]}{\partial F(\vec{x})} \right]_{F(\vec{x})=F_{k-1}(\vec{x})} \quad (8)$$

として求められる。ただしここで

$$F_{k-1}(\vec{x}) = \sum_{i=0}^{k-1} f_i(\vec{x}) \quad (9)$$

を用いた。 ρ_k については、 g_k を固定しながら、 f_k が最も大きく損失関数の期待値を下げるよう定め、

$$\rho_k = \arg \min_{\rho} E_{s, \vec{x}} (L(s, F_{k-1}(\vec{x})) - \rho g_k(\vec{x})) \quad (10)$$

となる。

こうして、差分 f_k を加えて F を更新する iteration を多数回繰り返し、収束を目指すことにより、トレーニングデータとして得られたスコア s を最も良く再現するであろう解 F^* を探索することができる。

4.3 スコアリングモデルの開発プロセス

Web 検索エンジンにおけるスコアリングモデルを準備するためには、他の多くの機械学習モデルと同様、図3のように学習アルゴリズム自体の開発、トレーニングデータ収集、モデルのトレーニング実施、テスト実施といった過程を経ることになる。また、relevanceの維持・改善のためには、この過程を一つのサイクルとして、継続的に繰り返し実施することが求められる。

Web 検索サービス運営において新たな機能をリリースする際には通常、品質に大きな問題がないことをオフラインテストで確認した後、ある限られた割合の

図3 スコアリングモデルの開発プロセス

ユーザへ向けてのテストリリースを実施し、クリック動向やクエリ総数の変動その他についてのモニタリングを行う。このようなオンラインテストはライブテストとも呼ばれ、ログの集計によってユーザ満足度とも関連する様々な統計情報を得ることができ、複数のセッティングがあり得る中でリリース対象の絞り込みや、最終的なリリース可否判断の根拠として用いられる。ランキングモデル・スコアリングモデルの開発においてもライブテストが実施され、その結果が最終的な品質評価に利用される。

ランキング・スコアリングモデルのリリースにあたっては当然ながら個別のWebページの検索結果上の順位に変動が見られることが多く、Yahoo! JAPANではランキングに関心のあるパブリッシャーへ向けたコミュニケーションのため、「アルゴリズムUpdate」として対外的なアナウンスを併せて実施している¹。

5. Web 検索にまつわるアドバンスドトピック

その他、Web 検索における比較的新しい問題領域から二点、興味深いトピックを取り上げる。一点目はrelevanceの評価において時間軸の考慮が必要なポイントであり、二点目は現在、普及が進みつつあるセマンティックウェブに対してWeb 検索サービスとしてどのように対応するか、どのような貢献を行いうるかに関連するものである。

5.1 周期性を持ったイベントについてのコンテンツ、それに対する検索要求

Web 検索のクエリにおいては「七夕」、「春節」、「大学入試」など、年周期で行われるようなイベントについて問い合わせられるものが一定数存在する。また、地域で行われる七夕関連の催事や、各大学の大学入試の案内など、そのようなイベントについてのコンテンツもWeb上に数多く提供されている。

ここで問題となるのは、例えば「隅田川花火大会」のように過去何度も開催されている花火大会についての検索が行われたとき、コンテンツとしては最新年のものを含め過去複数年分が存在する一方、検索者は通常は最新年のものだけを必要としているという点である。このような情報要求を満たすためには

- ・ 検索実行時に、クエリが年周期のイベントを指すものかどうか判定
- ・ 年周期のイベントを指すクエリに対して、最新年のコンテンツが検索結果として上位に供されるよ

1 <http://searchblog.yahoo.co.jp/>

う、検索結果のランキングをコントロール

という少なくとも二つの側面において対応を行わねばならない。後者のランキングの側面については、古い年度の実験結果が上位に現れるケースをエラーと捉えフィードバック制御を行うアプローチ[Zhang09]が既に検討されている。

5.2 セマンティックウェブへの対応

Web検索で扱うHTMLの情報は通常、自然文を中心としながらも、タイトル・見出し・段落などのタグ付けがなされ緩やかに構造化された形となっており、情報検索の分野ではしばしば半構造化データとして分類される[Manning08]。HTMLのタグセットは(metaタグのような例外はありながらも)本来的にはコンテンツの表示のされ方を規定するものであって、コンテンツ中の文字列に対して、例えば日付、価格、商品名といった「意味」をタグ付けするような役割は負っていない。しかしながら現在、RDFa、eRDF、Microformatなど、HTMLタグ中の文字列への様々な意味あるいは型の付与を可能とする、コンテンツのレイヤーにおけるセマンティックウェブ標準の普及が進んだことで、半構造化データであるHTMLの中に型付けされた構造化データを埋め込むことがしばしば行われるようになった。

このように提供された構造化データを、Web検索エンジンではどのように利用することができるのか。

回答の一つとしては、米Yahoo!により2008年6月に提供が開始されたSearchMonkeyのような機能が考えられる。Yahoo! Searchにおけるこの新機能は、RDFa、eRDF、MicroformatなどでWebページに埋め込まれた構造化データを、セマンティクスを失わない形で検索インデックス中に保持しておき、該当ページが検索結果にヒットした際、タイトル・サマリー・URLの組に留まらない、構造化されたリッチなフォーマットで検索結果を表示するものである。Googleもまた同様のアプローチによる新機能、Rich Snippetsを

2009年5月に公開している。

上述の一連の動きについては、コンテンツプロバイダーによるセマンティックデータ提供が検索サービス側の対応を促し、それがさらにコンテンツプロバイダーのデータ提供を加速し最終的にユーザにとっての利便性向上に繋がるような、良循環が始まっているものと理解することができよう。

日本国内向けにはGoogleのRich Snippetsが2009年11月から、Yahoo! JAPANのSearchMonkeyが「Yahoo! 検索プラグイン」の名称で同年12月からそれぞれ正式にサービスを開始しており、国内においてもWeb上のセマンティックデータ提供・利用のエコシステム実現が期待される。

また当然ながら、これらの構造化データを検索のrelevance向上にどのように生かしていくか、ということが今後のWeb検索サービス改善における大きなテーマとなるであろう。

6. おわりに

本稿では、現在の商用Web検索エンジンにおけるrelevanceを支えるアーキテクチャと、その評価方法について解説を行った。

転置インデックスによるテキスト全文検索の技術自体は既に確立されていると言っておそらく過言ではないが、しかしスマートフォンの普及や、Twitterなどのリアルタイムかつソーシャルな側面を持つ新しいサービスの利用拡大により、人々の生活における検索シーンや検索対象となりうるコンテンツは日々、その幅を拡大している。「クエリとドキュメントテキストの関連性」に留まらない、地理的な、時間的な、さらにはソーシャルな関連性が検索サービスに望まれており、それを実現するためのより一層の研究開発に向けて、本稿が一助となれば幸いである。

参考文献

- [Licklider68] J. C. R. Licklider : The Computer as a Communication Device, Science and Technology (1968)
- [Manning08] C. D. Manning, P. Raghavan and H. Schütze : Introduction to Information Retrieval : Cambridge University Press (2008)
- [Croft09] W. B. Croft, D. Metzler and T. Strohman : “Search Engines : Information Retrieval in Practice” : Addison Wesley (2009)
- [Jones72] K. Spärck Jones : A statistical interpretation of term specificity and its application in retrieval, Journal of Documentation 28 (1) pp. 11-21 (1972)
- [Kleinberg98] J. M. Kleinberg : Authoritative sources in a hyperlinked environment, Proceedings of ACM-SIAM Symposium on Discrete Algorithms (1998)

図4 Yahoo!検索 SearchMonkeyによる検索結果プレゼンテーションの改善

- [Brin98] S. Brin and L. Page : The anatomy of a large-scale hypertextual Web search engine, Proceedings of the 7th World Wide Web Conference (WWW7) pp.107-117 (1998)
- [Wikipedia10] Learning to rank, Wikipedia (2010) : http://en.wikipedia.org/wiki/Learning_to_rank
- [Friedman01] J. H. Friedman : "Greedy function approximation : A gradient boosting machine", Annals of Statistics 29, 5 pp.1189-1232 (2001)
- [Zhang09] R. Zhang, Y. Chang, Z. Zheng, D. Metzler and J. Nie : Search Result Re-ranking by Feedback Control Adjustment for Time-sensitive Query, Proceedings of

the North American Chapter of the Association for Computational Linguistics - Human Language Technologies (NAACL-HLT 2009)

(2010年2月19日 受付)

[問い合わせ先]

〒107-6211 東京都港区赤坂9-7-1 ミッドタウン・タワー
ヤフー株式会社 R&D 統括本部 プラットフォーム開発本部
清水 徹

TEL : 03-6864-3206

FAX : 03-6440-4300

E-mail : toshimiz@yahoo-corp.jp

著 者 紹 介

しみず とおる
清水 徹 [非会員]

1996年名古屋大学理学部物理学科卒業。1998年同大学理学研究科素粒子宇宙物理学専攻修士課程を修了。SAPジャパン(株)、(株)トライテック等を経て、2008年2月よりヤフー株式会社勤務。R&D統括本部 プラットフォーム開発本部においてWeb検索サービスのセマンティックウェブ対応、関連性改善および網羅性改善などの開発企画に従事。ACM SIGIR会員。