アクセス集中サイトを利用した検索精度の向上

検索エンジンはあらゆる情報の取得方法として欠かせない存在になっています．しかしながらWWW（World Wide Web）上の情報は爆発的に増加しており，ユーザの検索意図に沿った検索結果を返すことが困難になっています．本稿では次世代の検索エンジンに向けて，高精度な検索結果を実現するランキング技術について紹介します．

村田 眞哉（むらた まさや） /戸田 浩之（とだ ひろゆき）
松浦 由美子（まつうら ゆみこ） /片岡 良治（かたおか りょうじ）
NTTサイバーソリューション研究所

高精度検索エンジンに向けて

WWW上に存在する情報はもはや企業や個人のホームページのみならず，百科事典，論文，本，地図，ブログなど非常に多岐にわたり，あらゆる有益な情報が存在しています．それら情報の取得・閲覧手段として，検索エンジンは現在の情報社会を生きるうえで欠かせない存在になっています．

検索者（ユーザ）は検索キーワード（クエリ）を検索エンジンのインタフェースに入力するだけで，瞬時にそのクエリに関する検索結果の集合を取得することができ，さらにクエリとの関連性などに基づき順位付け（ランキング）された結果を閲覧することができます．

このランキングの精度を上げることにより，ユーザが膨大な情報の中でさまようことがなくなり，目的の情報へ効率的に到達することが可能となります．したがって，検索精度の向上においてランキングは重要な要素の１つであり，かつユーザにその有能性を印象付ける役割をも担っています．

今回紹介する技術は，今までの技術で達成している検索精度をさらに上回る，ユーザの検索意図を反映した高精度なランキングを実現します．それによりユーザにかける検索労力を最小化し，結果として多くのユーザに継続的に使っていただける検索エンジンをつくり上げることができます．

従来のランキング手法と提案手法

検索結果のランキングを決める要素として，クエリがサイト（本稿では個々のWebページのこと）の文章中にどれくらい含まれているか，サイトが他のどのようなサイトからどれだけリンクされているかなどがあり，それらをうまく組み合わせることで現在の高精度なランキングが実現されています．

それに加えて近年では検索結果があるクエリでどれだけクリックされているのかという，ユーザの検索意図を直接扱うことができる新たな要素が注目されています．この要素をうまく考慮することで従来のランキング手法ではとらえきれていなかった，ユーザの検索意図を反映したランキングが可能になります．

NTTグループでは，ポータルサイトgooにおいて各種検索エンジン[(1),(2)]を提供しており，実際に多くのユーザが検索機能を利用していることから，質の高い検索エンジンのログを得ることができます．我々はこのログに注目し，ログを解析することで得られるさまざまな情報を用いて検索精度を向上させるランキングについて研究を行っています．

検索エンジンのログ

検索エンジンのログには**図１**に示すように，ユーザが入力したクエリとそれに対して閲覧した検索結果の情報が残されています．以下ではユーザの検索結果の閲覧行動に関するログ（クリックログ）に関して説明します．

クリックログにはユーザが入力したクエリに対して実際にクリックした検索結果のURLとそのランキング，そしてその時刻が残されています．クリックログを解析することで，あるクエリに対してユーザがよくクリックしている検索結果を特定することができます．

このクリック回数はそれ自体有益な情報であり，かつランキングを決定する有効な要素の１つですが，クリックされた検索結果とされなかった検索結果にはどのような違いがあったかを考えることでさらに有益な情報を得ることができます．ユーザは検索結果のタイトルと概要文（スニペット）を見て，自分が求めている情報がその検索結果

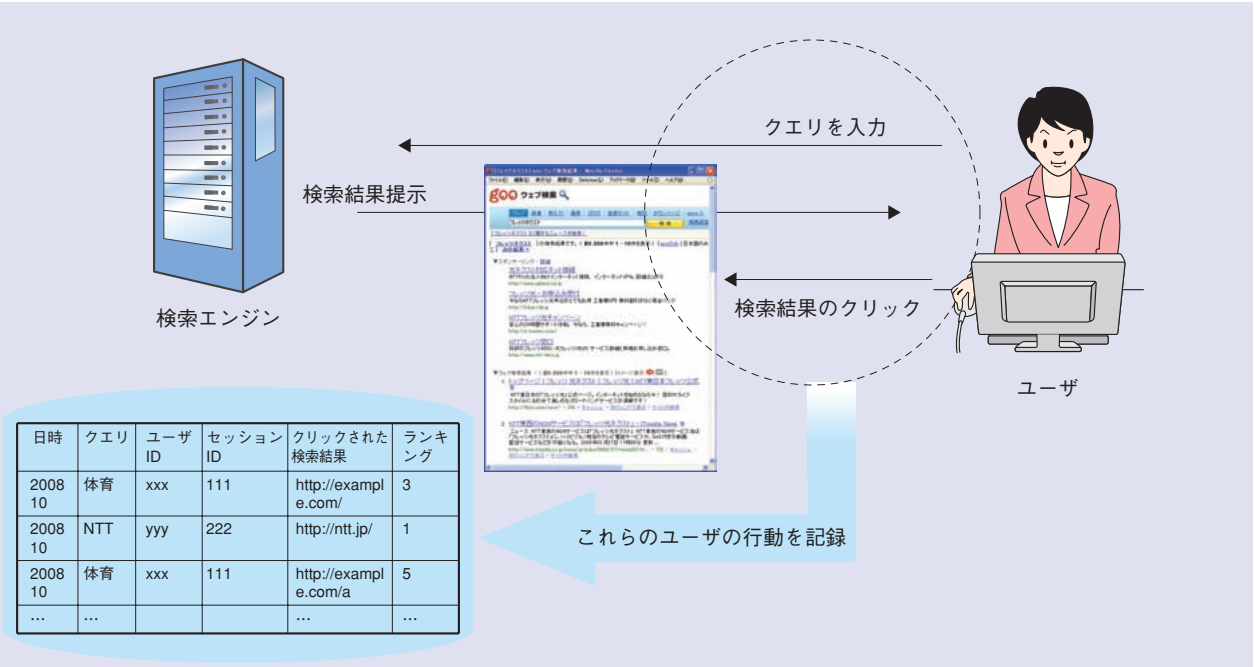


図1　検索エンジンのログ

の中に存在するかどうかを判断すると考えられるので，よくクリックされている検索結果のタイトルとスニペットにはユーザが入力したクエリの検索意図に関連する表現が含まれていたと考えられます．

実際このアイデアは，クリックログを利用した検索結果のランキングの基本的なアイデアとして幅広く採用されています[3]．我々の提案している手法もこのアイデアを基礎としています．

クエリ拡張法とは

検索エンジンのログを解析して得られた情報を，実際の検索結果のランキングに反映させる手法としてはクエリ拡張法（Query Expansion Method）と呼ばれるものが代表的です．クエリ拡張法についての説明を**図2**に示します．ユーザがクエリで表現した検索意図と関連が高いキーワード（拡張語）を取得し，このクエリに付与して検索を行うことで，ユーザが求める高精度なランキングを提示する手法です．

クエリ拡張の精度を決定付けるのは拡張語の質であり，現在では検索エンジンのログを利用して拡張語の取得を行う方法が研究されています．

代表的な手法としてCuiら[3]の研究があります．この手法で得られる拡張語はクエリの検索意図を表し，したがってこの拡張語を用いてクエリ拡張を行うことにより，クエリの検索意図に適合しているであろうサイトを網羅的にランキング上位に上げることができます．このような理由からクエリ拡張法がクリックログ解析の応用として幅広く用いられている手法になっています．

次項で紹介する我々の手法も，上述したクリックログを利用したクエリ拡張法の考え方を基礎に置いています．

アクセス集中サイトを利用したクエリ拡張法

ここから我々が提案している手法について説明します．前述したクリックログ解析を利用したクエリ拡張法の精度を最大限に高めるには，まず検索結果において真にクリックが集中しているサイトを特定することが必要になります．そしてそのようなサイトを特定することができれば，それらサイト群のタイトルとスニペットに共通して存在するキーワードを拡張語として抜き出し，この拡張語でクエリ拡張を行います．これはクエリの検索意図を補うことができる最良の拡張語でクエリ拡張を実行することを意味し，結果として検索精度の大幅な向上を望むことができます．我々はこの真にクリックが集中しているサイトをアクセス集中サイト（ACS: Access Concentration Site）と呼び，そのアクセスの集中度合をACD（Access Concentration Degree）と名付けました[4]．

Cuiら[3]の手法では，単純にクリック回数が多いサイトをACSとみなしています．しかしクリック回数にはランキング上位の検索結果ほどよくクリックされ，ランキング下位になるにつれて急激にクリック回数が減少するとい

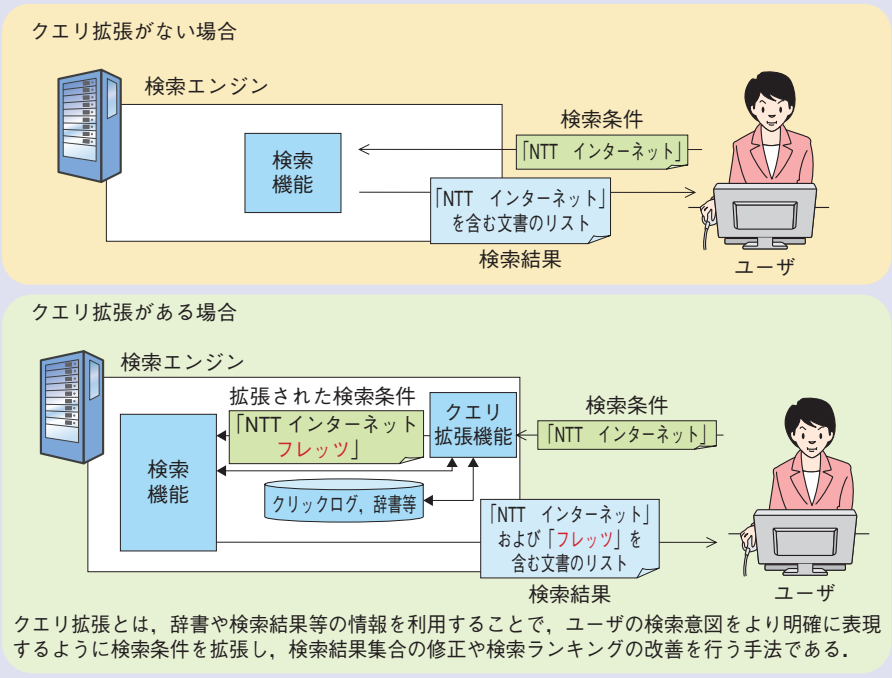


図 2　クエリ拡張法

うランキングによるバイアスが強くかかっています．これはユーザは検索結果の上位しか見ないという行為に起因するものです．したがって単純にクリック回数が多いサイトをACSとみなすことは，クリック回数が急激に少なくなる検索結果下位のサイトを無視し，上位のサイトのみ注目していることになります．

我々はこのクリック回数がランキングに強く依存するという問題に対処し，より上質なACSを特定するために以下のように考えました．**図 3** のアイデアの説明図を用いて説明します．ACSはその定義から分かるように非常に多くのクリック回数を獲得します．したがってあるクエリに対してクリックログを解析し，検索結果ランク（横軸）とクリック回数（縦軸）の座標上でこの解析結果の曲線（特定クエリクリック曲線）を描くと，ACSが存在する検索結果ランクにはピークが立つはずです．そしてこのピークの強さがACDに対応しています．この座標上において，あるサイトのクリック回数を頂点，その両隣のランクに存在するサイトのクリック回数を左右の頂点と見なすことにより，図 3 の右の図を描くことができます．するとピークの強さ，すなわちACDはこの図の頂点に対する左右の仰角の和で定義することができます．

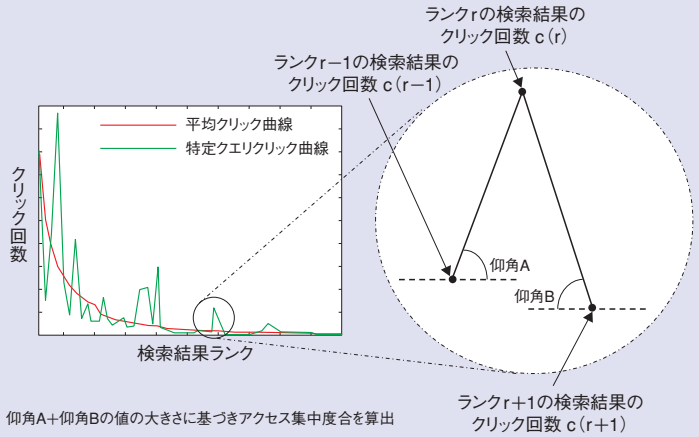
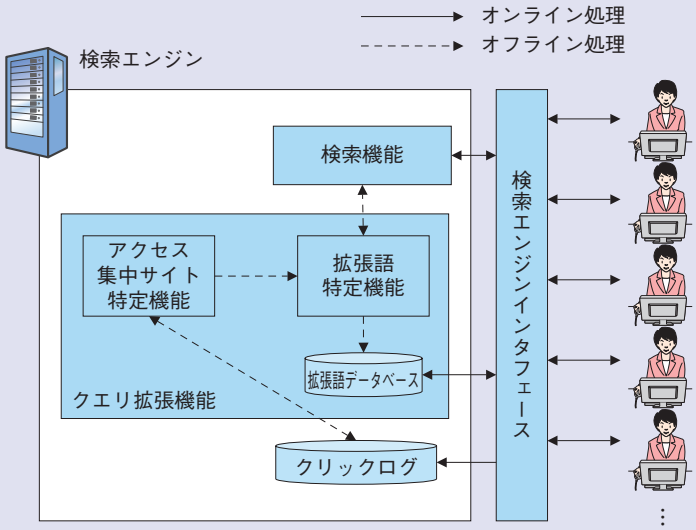


図 3　アクセス集中サイトの特定法

このACDの定義によりランキングのバイアスが軽減され，アクセスが集中しているサイトを特定することができます．この理由を図 3 を用いて説明します．ランキング r に存在するサイトのピークの強さはそのクリック回数 c(r) と両隣のランクのクリック回数 c(r−1)，c(r+1) により決定されます．一般的にランキング上位のサイトは下位のサイトと比べてよくクリック



オンライン処理
①ユーザから検索条件を受け付け
②検索条件を基に「拡張語データベース」に問い合わせ拡張語を取得
③拡張した検索条件で「検索機能」にアクセスし検索実行
④検索結果をユーザに提示
⑤ユーザが閲覧するために選択した情報を「クリックログ」として記録

オフライン処理
①「クリックログ」から各クエリごとのログを特定
②各クエリごとのログからアクセス集中サイト特定
③アクセス集中サイトの情報を基に，「検索機能」にアクセスし，ユーザが閲覧したタイトル，スニペットの情報を取得
④タイトル，スニペットの情報を基に拡張語を特定し，「拡張語データベース」に格納

図4　システム構成

されるので，c(r)が大きくなる傾向があります．しかしながらその両隣のランクのクリック回数c(r−1)，c(r+1)もこの傾向を受けて大きくなるので，我々が定義したACD，つまりピークの計算ではこのc(r−1)，c(r+1)がランキングのバイアスを打ち消します．その結果，真にアクセスが集中しているサイトを特定することが可能になります．クリックログを解析して全サイトのACDを計算し，このACDの高い上位のサイトを我々はACSと定義します．

あるクエリに対するACSが特定されると，それらのタイトルとスニペットからキーワードを抜き出します．このキーワードは各ACSに共通して存在し，かつ一般的に稀であるものから優先的に抽出していきます．これによりユーザがクエリで表現した検索意図と関連が高いキーワードを取得することが可能となります．そしてこのキーワードでクエリ拡張を行うことにより，ユーザが求めている検索結果をランキング上位に表示することが可能となり，検索にかける労力を減らすことができます．

システム概要

システム構成を**図4**に示します．各クエリに対する拡張語の取得をオフライン処理で行い，クエリ拡張の処理がオンライン処理で行われます．

今後の展開

本稿ではアクセス集中サイトを利用した検索精度の向上に関する技術を紹介しました．今後も高精度な検索サービスの実現に向け，次世代の情報検索技術に関する研究を進めていきます．

■参考文献

(1) http://www.goo.ne.jp/
(2) http://mobile.goo.ne.jp/
(3) H. Cui, J.R.Wen, J.Y. Nie, and W.Y.Ma：“Probabilistic Query Expansion Using Query Logs，” Proc. of WWW'02, pp.325-332, 2002.
(4) M. Murata, H. Toda, Y. Matsuura, and R. Kataoka：“Improving Mobile Web-IR Using Access Concentration Sites in Search Results，” Proc. of WISE'08, pp.221-234, 2008.

（上段左から）戸田　浩之/ 村田　眞哉
（下段左から）片岡　良治/ 松浦　由美子

高精度な検索サービスの実現に向け，引き続き研究開発を進めていきます．

◆問い合わせ先
NTTサイバーソリューション研究所
メディアコンピューティングプロジェクト
TEL　046-859-3394
FAX　046-855-1730
E-mail　murata.masaya@lab.ntt.co.jp