1G1-1

QueReSeek: 検索履歴共有によるコミュニティ指向の連想検索

QueReSeek: Community-Oriented Associative Search by Sharing of Search History

丹 英之[*1] 大向 一輝[*2] 武田 英明[*2]
Hideyuki TAN Ikki OHMUKAI Hideaki TAKEDA

[*1]株式会社アルファシステムズ [*2]国立情報学研究所
Alpha Systems Inc. National Institute of Informatics

In this paper, we propose a method of associative search. This method can obtain the useful retrieval results for the community by sharing of search history. The users who have a common purpose often search and browse similar Web contents. This method presents the user contents of neighborhood in Web contents under browsing by using search query that collects with the Web Proxy. As a result, the user gets the search result that the concern of the community is reflected. It is thought that the bottom rising of the retrieval ability by the entire user who belongs to the community can be expected, and the knowledge sharing is promoted.

1. はじめに

本稿のようなコンテンツをウェブ経由で入手し，PC のモニタで閲覧するのが当然のようになった昨今，“知りたい事があったらまずインターネットの検索エンジンで” という情報探索行動は，ウェブを利用したことのあるユーザなら誰もが取るようになった．しかし，ウェブログをはじめとした一般の利用者自身がコンテンツを生成するメディアの台頭もあり，インターネット上で生成・流通されるコンテンツは日々増加し，知りたい事柄についての情報・知識を簡単に入手することは困難になってきた．

そこで我々は，ある共通の目的を持った検索エンジンユーザの集団をコミュニティと定義し，このコミュニティの力を借りることで，目的とする情報・知識が掲載されているウェブコンテンツへ閲覧を誘導する手法を検討してきた [丹 07]．ここでいう共通の目的とは，同じ領域に興味・関心を持っており，ある程度共有された知識ドメインをテーマとして検索エンジンを利用している，ということである．また，コミュニティは似たような境遇・環境下にいる人々の集まりで，例えば会社や学校の研究室などを単位とした集団が対象になる．すると，このコミュニティを形成するユーザが検索エンジンに発行する検索クエリは，コミュニティを形づくる共通の目的と潜在的に関連を持つであろうと想定できる．つまり，コミュニティ内から発生した検索クエリは，潜在的にコミュニティの目的のためであり，コミュニティ内にて関心のあるウェブコンテンツを探し出すことができる．この検索クエリを再利用することで，閲覧中のウェブコンテンツに対し，そのウェブコンテンツがどのような検索クエリによってこのウェブコンテンツに辿り着くことができるかを提示する．この手法は，ある共通の目的を持つユーザ同士では，似た内容のウェブコンテンツを検索・閲覧している場合が多い [丹 06]，という考えに成り立っている．

我々はこの手法を基に，閲覧中のウェブコンテンツに対し検索クエリを提示するだけでなく，更に関連するウェブコンテンツをユーザへ提示できるよう拡張した．この提示する関連コンテンツは，閲覧中のコンテンツとコミュニティが利用した検索クエリで繋がる，閲覧中のウェブコンテンツと内容的に近傍にあるウェブコンテンツである．これにより，ユーザはコミュニティの関心を反映した連想検索の結果を得ることができる．よって，ユーザは効率よく情報探索を行うことができるようになり，コミュニティに属するユーザ全体での検索能力の底上げが期待されるとともに，延いてはコミュニティ内にて知識共有が促進されると考えられる．

本稿では，ユーザの属するコミュニティ内にて検索履歴を共有することで実現される，コミュニティに効果的なウェブコンテンツの連想検索方法を提案し，この提案手法の評価について述べる．

連絡先: 丹英之，株式会社アルファシステムズ，川崎市中原区上小田中 6-6-1，044-733-4111，tanh@alpha.co.jp

2. 関連研究

以下に，検索クエリとウェブページの関係，言葉による繋がりを用いたコンテンツ検索である連想検索，そして，既に商業サービスが行われているインターネット検索エンジンの類似ページ検索を本研究の関連研究として紹介する．

2.1 検索クエリとウェブページの関係

クエリログからウェブページの典型的クエリを抽出する試みがある [甲谷 07]．本提案手法の考え方と同様，典型的クエリとは潜在的に存在するコンテンツ利用者のニーズを端的に表す要約であるとする点が類似しており，検索結果からの閲覧遷移を用いた重み付けによってクエリログから抽出した検索クエリ候補を絞り込む．この手法では，ユーザの集合としてのコミュニティについては触れておらず，典型的且つ汎用的な検索クエリを抽出することを目指しており，ある目的を共有しているコミュニティに対して価値のある検索クエリを抽出することは企図していない．

本提案手法は，共通の目的のもとに集ったユーザをコミュニティとし，その背景にある知識を表すものとして検索クエリを共有し再利用する手法である．

2.2 連想検索

ドキュメント中にある言葉の重なり度合によってドキュメント間の距離を算出し，それを元に類似するドキュメントを検索する．この先行研究に，汎用連想検索エンジン GETA(Generic Engine for Transposable Association) を挙げる [高野 02]．GETA では，文書-単語行列を元に TF,IDF など各種の統計的類似性計算を行うことで文書間，単語間の関係を求めることができる．GETA を用いた連想検索では，検索クエリとして

与えられた単語の近傍にある言葉をドキュメント集合から取り出す．つまり特徴語で繋がったドキュメント集合を検索結果として与える．

一方，本提案の手法では，すべてコミュニティが生成した検索クエリを近傍にある言葉として利用するので，ウェブ空間からコミュニティにとって重要な言葉で繋がったウェブコンテンツを検索結果として取り出す．よって，コミュニティにとって有用なウェブコンテンツを検索できる．

2.3 インターネット検索エンジンによる類似ページ検索

インターネット検索エンジンサービスである Google は特殊サーチの機能として指定したウェブコンテンツに似ているページを検索する類似ページ検索[*1] を提供している．これは関連ページ検索とも呼ばれ，“related:” の接頭辞をつけた URL を検索クエリとして受け取り，関連するウェブコンテンツの一覧をユーザへ返す．関連ページを導き出すアルゴリズムは非公開であるが，ウェブコンテンツに含まれるアンカーテキストやリンク先ページ，リンク元ページなど周辺にあるリンク構造を用いた手法であると推測される．

ユーザからの入力である検索クエリが検索の起点となるウェブコンテンツを指す URL である点，及び，ユーザへの出力が関連するウェブコンテンツ集合を指す URL である点が本提案手法と同じである．つまり，ユーザとシステムのインタフェースに互換性がある．

3. 提案手法

ここでは提案する手法について，検索クエリで繋がるウェブコンテンツ，そして，閲覧中のウェブコンテンツに因む検索結果について述べる．

3.1 検索クエリで繋がるウェブコンテンツ

検索エンジンを利用するユーザは，ある事柄に関する情報・知識を得たいと考えているはずである．そのため，ユーザが検索エンジンに対して発行する検索クエリは，ユーザの持つ背景知識を基にして生成され，望んでいるウェブコンテンツに関連するであろうという思慮を含んだ文字列になる．そして，検索エンジンはロボットによって収集したウェブコンテンツ集合に対して全文検索を行い，採用しているアルゴリズムによってスコアを算出することで順位を決定し，結果としてウェブコンテンツの URL 一覧をユーザへ返す．検索クエリがあって，検索結果集合ができる．つまり，検索クエリはウェブコンテンツを対象とした全文検索の結果集合に含まれる文字列であり，ウェブコンテンツの断片としてランキングアルゴリズムによってウェブコンテンツと結びついている．検索クエリと検索結果として得られるウェブコンテンツを指す URL 集合の二部グラフを図 1 に示す．検索クエリ A の検索結果として得られるウェブコンテンツを指す URL を，結果のランキング順に url_{A1}, url_{A2}, ... と表している．この関係を逆から見ると，ウェブコンテンツ url_{A1}, url_{A2}, ... が検索クエリ A の一言になる．つまり，ウェブコンテンツが一言に抽象化されたことに相当する．提案手法は，検索クエリとそれを用いた検索で得られたウェブコンテンツとの間にある関係，“ウェブコンテンツを一言で表すと，検索クエリになる” という関係を用いる．

ここで，検索クエリ B の検索結果である URL の集合も同様に示すと，検索クエリの近さによっては，同じウェブコンテンツを指す URL が得られる場合がある．図では url_{B1}, url_{B2}, url_{B3} がそれぞれ，url_{A2}, url_{A3}, url_{A4} と同じウェブコンテンツを指している．ここで，$\mathrm{url}_{A2}(\mathrm{url}_{B1})$ が閲覧中のウェブコンテンツであるとするならば，$\mathrm{url}_{B2}(\mathrm{url}_{A3})$, $\mathrm{url}_{B3}(\mathrm{url}_{A4})$ は，検索クエリ A, B で繋がったウェブコンテンツになる．そして検索クエリ A, B は，コミュニティが生成したものなので，$\mathrm{url}_{B2}(\mathrm{url}_{A3})$ と $\mathrm{url}_{B3}(\mathrm{url}_{A4})$ は，コミュニティの興味・関心を反映したウェブコンテンツということになる．これらの URL をユーザへ提示する．

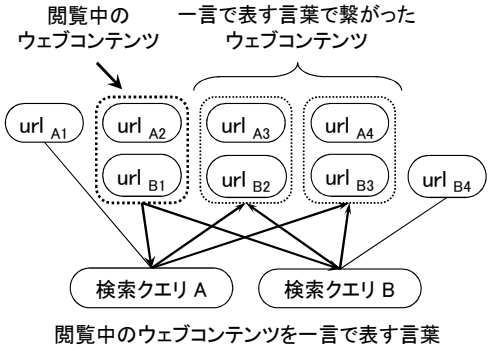

図 1: 検索クエリ-検索結果 URL 集合の二部グラフ

検索エンジンを利用した情報探索行動において，一回の検索でユーザの望むウェブコンテンツに辿り着かない場合，ユーザは検索クエリの文字列を替え再び検索を行う．そして，検索結果で得られたウェブコンテンツの確認や検索クエリの変更など試行錯誤を，望むウェブコンテンツに辿り着くまで繰り返すことになる．この試行錯誤の過程において，ウェブコンテンツの俳徊的な閲覧行動に伴い，ユーザの持っている知識は変化していく．そして，知りたい事柄，つまり未知であった事柄に関連する知識が増えることでユーザの生成する検索クエリは，ユーザの望む情報・知識へと近づいていく．知りたい事柄が記載されているウェブコンテンツへ辿り着いた暁には，得られたウェブコンテンツの内容とユーザの知りたかった事柄，そして編み出した検索クエリが強固に結びつく．この，ユーザが望んでいる情報・知識へと少しずつ近づいていく過程の中で，ユーザの生成する検索クエリは，知りたい事柄に対する絞り込みのフィルタとして機能する．つまり，検索クエリは情報探索目的への手がかり，言い換えるとユーザの情報探索ノウハウが形式化されたものであると言える．

先に述べたとおり，ある共通の目的を持つユーザ同士では，似た内容のウェブコンテンツを探して閲覧している場合が多い．すると，このユーザ同士が属するコミュニティにて検索クエリを共有することで，形式化された情報探索ノウハウが共有できることになる．この，コミュニティが利用した検索クエリをユーザへ還元する手法が [丹 07] である．本稿で提案する手法は，閲覧中のウェブコンテンツを基に，形式化された情報検索ノウハウによって結び付いたウェブコンテンツをユーザへ提示する．

3.2 閲覧中のウェブコンテンツに因む検索結果

複数のウェブコンテンツをユーザへ提示するには，それらを提示する順番を決める必要がある．検索クエリと検索結果である URL 集合のセットを大量に用意しておくことで，閲覧中のウェブコンテンツが検索結果として得られる検索クエリを複数得ることができる．この得られた検索クエリをコミュニティで利用された回数，検索結果のヒット数，検索結果集合内

*1 Google の特殊機能 関連ページ
http://www.google.com/help/features.html#related

における閲覧中ウェブコンテンツの順位，の順でソートする．すると，コミュニティの関心・興味を反映する検索クエリ集合が上位にくるので，これらを繋がりとして利用する．繋がりとなる検索クエリが判ると，関連するウェブコンテンツを得ることができる．これには，検索エンジンから得られた検索結果集合内での順位を用い，この順位の逆数を積算したものをスコアとし，このスコアが高いものから順に提示する．これらの処理により，閲覧中のウェブコンテンツに対し，コミュニティの関心・興味を反映した関連度の高いウェブコンテンツを得ることができる．

4. 評価

ここでは，提案手法の有効性を検証するために行った評価実験について述べる．提案手法を実装したシステムは，ウェブコンテンツを検索クエリとした検索エンジンに他ならない．これは，関連研究で紹介した Google の特殊クエリを用いた類似ページ検索と同じである．そこで，Google の類似ページ検索 (以下，related 検索と呼ぶ) で得られた結果との比較による評価実験を行うことにした．

4.1 実験準備

まず，コミュニティの利用した検索クエリを収集する必要がある．これには，筆者の属する組織の内，約 1,500 ユーザ，4 か月のウェブ閲覧履歴が記録されている Web Proxy のログを対象に，検索エンジン Google[*2], Yahoo![*3], Live Search[*4], Excite[*5], goo[*6], Infoseek[*7] へ投入された，利用頻度の高い検索クエリ 50,762 個を抽出した．この抽出した検索クエリを，Googe, Yahoo!, Live Search 三つの検索エンジンで再検索し，延べ 7,352,869 個，ユニークでは 4,632,628 個の URL を収集した．検索クエリと検索結果の URL は，評価実験の期間中にも被験者の検索エンジンの利用を検出する度に追加していく．

実験には，筆者と同じ組織に属する検索エンジン利用歴が 5 年以上の 4 人が被験者として参加した．これら被験者には提案手法を実装したシステムについての説明を行った．

4.2 評価手順

一般に情報検索における被験者実験には，正解セットのある課題を被験者に与え，その課題の解決する様を観察する．今回は，業務中における実際の効果を観察するため，被験者が通常のウェブ閲覧において検索エンジンを利用した際の目的を，そのまま提案手法を評価するための検索課題として扱うことにした．但し，検索結果のリンクを一回辿ることで解決する容易な探索や，ナビゲーショナルクエリ [Broder 02] を用いた検索は除外した．

被験者は，業務中の検索エンジンの利用を伴ったウェブ閲覧行動において，課題設定に相当する状況下になったと判断した際，ブックマークレットを用い好きなタイミングで評価システムを呼び出す．すると評価システムは，提案手法と related 検索の結果が存在することを確認し，図 2 の画面を提示する．画面中，左右のカラムには提案手法及び related 検索の結果が，それぞれ上位 10 個並ぶ．左右のカラム，どちらが提案手法と related 検索の結果であるかはランダムで決定した．これら結果には，ウェブコンテンツのタイトル，URL，そして設問と

*2 http://www.google.co{m,.co.jp}
*3 http://www.yahoo.co{m,.co.jp}
*4 http://www.live.com
*5 http://www.excite.co.jp
*6 http://www.goo.ne.jp
*7 http://www.infoseek.co.jp

図 2: 提示ウェブコンテンツの評価用画面

その答えを入力するラジオボタンが一緒に提示される．被験者は，ウェブコンテンツのタイトルをクリックし，新しいウィンドウで該当するウェブコンテンツを確認し，設問に答える．コンテンツの評価には，Q1「Web 閲覧の目的に一致していたコンテンツですか?」と Q2「興味を惹いたコンテンツですか?」の二つの設問，そしてこれらの回答には,「はい」，どちらかと言えば「はい」，どちらかと言えば「いいえ」,「いいえ」の 4 段階の答えを用意した．

被験者は，この手順を検索の目的に合わせて 5 回行う．よって，被験者は合計 100 個のウェブコンテンツについて設問に答えることとなる．

4.3 実験結果

実験の間，評価システムは 130 回呼び出された．この内，提案手法による検索結果が得られなかったのが 22 回，Google の rerated:検索による結果が得られなかったのが 9 回，両方の検索結果が得られなかったのが 17 回であった．つまり，被験者の評価システム呼び出しの約 63%に対し評価可能な状態を提供できたことになる．また，追加された検索クエリは 627 個で追加分の延べ URL 数は，163,294 であった．

1 回のシステム評価で得られる，提案手法，related 検索それぞれの検索結果として得られたウェブコンテンツに対する評価回答 10 個分の平均値を，4 人 5 回分，まとめてプロットしたものを図 3 に示す.「Web 閲覧の目的に一致していたコンテンツですか?」と「興味を惹いたコンテンツですか?」の設問は，グラフのラベル「ウェブ閲覧の目的一致度」と「ウェブコンテンツへの興味度」に相当する．

提案手法では，ウェブ閲覧の目的一致度と興味度に相関がみられる傾向にある結果を得た．一方，related 検索では，目的一致度も興味度も低いコンテンツを提示した結果となった．

5. 考察

図 3 のグラフの第一象限に分類されたコンテンツは，ウェブ閲覧の目的に一致，つまり被験者のウェブ閲覧の文脈に沿っていて，且つ，被験者の興味を惹いたコンテンツである．よって，被験者にとって有用なウェブコンテンツを提示できたことになる．一方，第三象限はウェブ閲覧の目的に不一致で，更に興味を惹かないコンテンツであるので，被験者にとって無益なウェブコンテンツを提示したことになる．第二象限には，ウェブ閲覧の文脈に沿ってはいるが，興味を持てないウェブコンテンツが分類される．これは，既知の内容が掲載されたウェブコンテンツであると言えよう．一方，第四象限には，ウェブ閲覧の文脈に一致していないが，興味深いコンテンツが分類される．これは，その時の被験者の目的達成には相応しくないが，

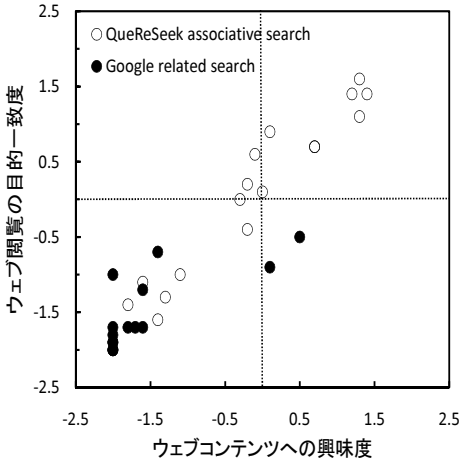

図 3: 提案手法と Google の類似ページ検索によって得られたウェブコンテンツの評価結果

関心を持ったウェブコンテンツを提示できたことになる．つまり，被験者へ未知のコンテンツとの遭遇を提供できたことになると言える．

この分類で得られた結果を検討すると，提案手法は第一象限から第三象限にわたって分類されており，提示したウェブコンテンツのうち，半分は正解であったが，もう半分は不要な提示であったことになる．本提案手法によって得られた検索結果は被験者にとって有用から無用まで幅広い評価を得るウェブコンテンツを提示した．一方，related 検索では，殆どが第三象限に分類された．今回用意した二つの設問における視点からは，提案手法に比べ悪い評価となるウェブコンテンツを提示した．related 検索で用いられているアルゴリズムは非公開なので推測の範囲での議論になるが，これは二つの手法のスコアリングの違いによるものであると考えられる．本提案手法ではウェブコンテンツに含まれる単語の繋がりによる内容に基づいたスコアリングによる検索結果を提示しているので，ウェブ閲覧の文脈に一致している感じる場合が多い．一方，related 検索では内容よりもリンク構造に重みを置いたスコアリングであるので，ウェブ閲覧の文脈に沿ったウェブコンテンツでないと判断されたと考えられる．この related 検索によって得られた文脈に沿っていないウェブコンテンツのうち，被験者の興味を惹いたものが若干ある．これが第四象限に分類されたウェブコンテンツであり，被験者へ未知のウェブコンテンツとの出会いを提示できたことになる．新しい出会いの提示が成功するかは，その提示されたウェブコンテンツに被験者が興味を持つかどうかにかかっている．related 検索では，この興味を惹くウェブコンテンツを十分ではないが，提示できた．

ところが提案手法では，被験者の興味を惹いたウェブコンテンツを提示できたものの，その提示ではウェブ閲覧の目的に一致したと判断されたものが多く，少し閲覧の目的から外れたウェブコンテンツは提示できなかった．これは，先に述べたスコアリングが内容によるものであり，利用頻度の高い検索クエリの影響が強かったとも考えられる．ユーザの携わる業務によっては，情報探索効率を高めるような第一象限に分類されるウェブコンテンツだけを提示するよりも，新しい着想を生み出すヒントとなるような第四象限に分類されるウェブコンテンツも含めて提示した方が，情報検索行為の補助として効果的であることも考えられる．この点を踏まえると，スコアリングの方法には，まだ十分に検討の余地があると言える．

6. まとめと今後の課題

本稿では，検索履歴を共有することで，ユーザの属するコミュニティに特化した検索結果を与える連想検索方法を提案し，得られる検索結果の評価について議論した．

提案した手法は，閲覧中のウェブコンテンツに対し，コミュニティが検索エンジンへ投入した検索クエリ，つまりコミュニティ内の重要単語によって繋がった内容的に近傍のウェブコンテンツをユーザへ提示する．これにより，ユーザはコミュニティの興味・関心を反映した検索結果を得ることができる．この手法の評価実験では，得られる検索結果のうち，ユーザのウェブ閲覧の目的に一致し，且つユーザの興味を惹いたウェブコンテンツであったものが約半分であった．

本提案手法では，利用されるに従って検索クエリと検索結果 URL 一覧のエントリが増加する．検索エンジンサイトでは，ロボットによる定期的なクロールにより検索結果の鮮度を保っている．特に，Google では “QDF”(Query Deserves Freshness) と呼ばれるアルゴリズムにより，検索を行った時期によって検索結果が大きく変動する．これらを考慮すると，検索クエリと検索結果 URL 一覧のエントリも定期的にアップデートする必要があるであろう．

また，コミュニティから回収した検索クエリには，検索結果ヒット数がゼロになるものや，単なる入力ミスで意味を成さない文字列など，塵芥の検索クエリも存在する．これらは，利用頻度や利用時期などによるフィルタリングで取り除くことが可能である．そして，ウェブコンテンツと検索クエリ間にはある程度関連があり知識ドメインが絞られるので，提案手法による結果は有用でない検索クエリに対し，ロバスト性を持つと考えられる．例えば休憩時間における私的な情報探索などにより，コミュニティの目的にそぐわない検索クエリが増加しても，得られる結果に害を及ぼす可能性は低い．

今後は，検索結果のスコアリングアルゴリズムの検討を進め，本稿の提案手法とこれまで検討してきたコミュニティ内の検索クエリ共有逆引きエンジンと組み合わせ，コミュニティと検索エンジンのコラボレーションによるウェブコンテンツのメタデータ生成プラットフォーム QueReSeek の構築を行う．

参考文献

[丹 07] 丹, 大向, 武田:QueReSeek:検索履歴の逆引きによるコミュニティベースの Web ナビゲーション, 人工知能学会全国大会 (第 21 回) 論文集 (2007).

[丹 06] 丹, 本田, 芝崎, 山口, 千葉, 原:Proxy で抽出した組織内ユーザの Web 閲覧特徴の時系列変化, 人工知能学会全国大会 (第 20 回) 論文集 (2006)

[甲谷 07] 甲谷, 湯本, 小山, 田中:Web ページに対する典型的なクエリの発見, 夏のデータベースワークショップ 2007, 情報処理学会研究報告 データベース・システム研究会報告 (2007)

[高野 02] 高野, 西岡, 今, 岩山, 丹羽, 久光, 藤尾, 徳永, 奥村, 望月, 野本: 汎用連想検索エンジンの開発と大規模文書分析への応用, 情報処理振興事業協会「独創的情報技術育成事業」2001 年度成果報告論文 (2002).

[Broder 02] A. Broder:A taxonomy of web search, ACM SIGIR Forum, Vol. 36, No. 2 (2007).