RD-001

クリックスルーに基づく探検型検索サイトの設計と開発
Design and Development of an Exploratory Search System based on Clickthroughs

酒井 哲也　小山田 浩史　野上 謙一　北村 仁美　梶浦 正浩　東 美奈子
野中 由美子　小野 雅也　菊池 豊
Tetsuya Sakai, Hiroshi Oyamada, Kenichi Nogami, Hitomi Kitamura, Masahiro Kajiura,
Minako Higashi, Yumiko Nonaka, Masaya Ono and Yutaka Kikuchi
(株) ニューズウォッチ
NewsWatch, Inc.

1. はじめに

近年，ユーザにより形式化された要求 (formalized need) もしくは調整された要求 (compromised need)[6] の存在を前提とし，ランクつき検索結果を出力して終了する従来型情報検索の枠組みを超えるものとして，探検型検索 (exploratory search) * が注目を集めている [8, 9]. これは，ユーザの情報要求が明確化するまでの過程と，ユーザが最終的な情報に辿りつくまでの過程を包含した広義の情報検索を意味し，特にユーザの情報要求の変化に対応することの重要性は 1980 年代から認識されていたが [1]，その評価方法論は未確立である.

(株) ニューズウォッチでは，ユーザの情報要求の変化をシステム側が促進し，繰り返し検索を実行してもらうことにより，ユーザが予期しなかった有用な情報を提供する探検型検索サイト「コトバ ノ ウチュウ」†の構築を進めている.「コトバ ノ ウチュウ」では，情報要求の変化を誘発するために，日本語版 Wikipedia の参照関係を視覚化したインタフェースを提供している‡. これに先立ち，我々は 2007 年 11 月にプロトタイプ検索サイトを Web 公開し，2008 年 2 月までの 4ヶ月間にわたりクエリログおよびクリックスルーを収集した. 本稿では，収集したデータに基づく評価結果および知見について，およびこれらをどのように「コトバ ノ ウチュウ」の設計と開発に活かしたかについて報告する.

2. 従来研究

探検型検索におけるユーザの情報要求の変化を扱う試みは「連想検索」とも呼ばれ，いくつかのシステムが既に Web 上で公開されている [2, 5]. 本研究は次章で示すように，少なくとも日本語版 Wikipedia の参照関係を視覚化し，現在のクエリに相当する Wikipedia エントリとここから参照されている他の Wikipedia エントリの間の関係を「ちら見せ」することによりユーザの興味の変化を積極的に促すインタフェースを備えている点で，これらの「連想検索」システムとは異なる.

Wikipedia の参照関係を扱う研究は，参照関係の種別を人手により記述する試み [7] と，これらを自動抽出する試みに大別される. 後者には，例えば Wikipedia エントリ間の "is-a" 関係などを自動抽出しオントロジ構築を試みた中山ら [4] らの研究がある. 本研究も Wikipedia の参照関係を自動抽出するものではあるが，研究の主眼は抽出方式自体ではなく，参照関係の「ちら見せ」により如何にユーザを新たな有用な情報まで導けるかにある. なお，我々の「ちら見せ」とは，現在のクエリに相当するエントリ内における他のエントリへの参照箇所の前後の文脈テキストをヒューリスティクスにより切り出して提示する比較的単純なものである.

本研究では，ユーザの情報要求の変化を誘発させる機能の改善を目的とし，プロトタイプ検索サイトのクエリログおよびクリックスルーを分析しているが，近年，クリックスルーを implicit relevance feedback に応用する研究が盛んである. 例えば，Joachims ら [3] は Web 検索ユーザが上位の文書をクリックしやすいという分析結果を報告している. 今回報告する分析結果の一部は，この「ランク重視」の傾向が検索対象の異なる複数の検索結果を同時に提示するインタフェースにも当てはまることを示唆している.

3. プロトタイプ

我々は，日本語版 Wikipedia の参照関係を利用してユーザの情報要求の変化を誘発させるインタフェースに対するユーザの反応を観察するため，2007 年 11 月 1 日にプロトタイプ検索サイトを Web ポータル「フレッシュアイ」§ 上に試験公開し，11 月 20 日にプレスリリースを行った. 12 月 11 日には，Yahoo!ニュース¶の検索結果画面から上記プロトタイプへの検索リンクが設けられた. このプロトタイプは，図 1 のようなレイアウトで，ニュース・Web・Wikipedia の各検索結果に加え Wikipedia を視覚化したクリッカブルなインタフェース「つながルート」をユーザに提示する.「つながルート」は Wikipedia の人名エントリのみを対象としており，検索クエリが人名エントリ名にマッチした場合にのみ提示される. その中心ノードは現在のクエリを表しており，周辺ノードは，現在のクエリに相当する Wikipedia エントリの中で参照されている他の人名エントリを表す. 周辺ノードにマウスオーバーすると，中心ノードに相当するエントリの中の該当参照箇所の前後の文脈テキストが表示される. また，周辺ノードをマウスクリックすると，周辺ノードの人名をクエリとした新たな検索結果が生成され，上記周辺ノードは「つながルート」内で中心ノードの位置に移動する. もしユーザが参照関係に興味を持ち続ければ，周辺ノードを辿り続け，検索開始時には考えてもみなかっ

*探索型検索と訳されることが多いが，探索も検索も本来 search の訳語であるため，本稿では探検型検索と呼ぶ.
†http://kotochu.fresheye.com/
‡(株) ニューズウォッチは，Wikimedia Foundation, Inc. との契約に基づき Wikipedia のコンテンツを利用している.
§http://www.fresheye.com/
¶http://headlines.yahoo.co.jp/hl/

図 1: つながルートを含むプロトタイプ検索結果画面

た情報を取得できる可能性がある．一方，サイト運営者としては，この仕組みによるページビュー増が利益に直結する．

図 2 および 3 に，2007 年 11 月～2008 年 2 月の 4ヶ月間におけるプロトタイプのページビューおよびクリック数を示す．まず，Yahoo!ニュースからの検索リンクの設置によりページビュー・クリック数共に大幅に増加している様子がわかる．図 3 は，図 1 に示した画面中のどの箇所がユーザにより多くクリックされたかを示している．Yahoo!ニュースから検索リンクが設置される以前 (これを期間 1 と呼ぶ) は「つながルート」が最も多くクリックされていたが，検索リンクが設置された後 (これを期間 2 と呼ぶ) は Web 検索結果の 1 件目が最も多くクリックされていたことがわかる．

表 1 は図 3 のクリック数を期間 1 と期間 2 に分けて一日平均を算出し，これに基づきソートしたものである．ただし，プレスリリースを行った 11 月 21 日のデータは期間 1 における異常値と見なし除外している．この表からも，期間 1 においては「つながルート」のクリック回数が非常に多かったのに対し，期間 2 においてはクリックされる割合が激減し，変わりに Web 検索がトップに来ていることがわかる．これは，Yahoo!ニュースから流入したクエリ中には人名クエリが極めて少なく，プロトタイプでは人名以外のクエリに対してそもそも「つながルート」を表示できていなかったことが原因であった．具体的には，Yahoo!ニュースから流入した異なり数 107,263 のクエリと，Wikipedia 人名エントリより作成した 105,980 件の「つながルート」の重なりは，わずか 2,844 件であった．従って，もし仮に期間 2 においても多くの場合に「つながルート」を提示できていたならば，ユーザはおそらく個々の検索結果よりも「つながルート」を好んでクリックしたものと推測できる．

一方，各検索結果のクリック数を見てみると，単一の検索結果を提示した場合の Joachims ら [3] の結果と同様，ユーザは検索結果の 3 位のページよりも 2 位のページを，2 位のページよりも 1 位のページをクリックする傾向があることがわかる．図 1 のようなレイアウトであっ

表 1: プロトタイプ検索のクリック数平均

期間 1(11/01-12/10，ただし 11/21 を除く 39 日間) の平均)		
つながルート	**52.6**	41.8%
ウィキペディア 1 件目	17.1	13.6%
ウェブ 1 件目	12.7	10.1%
ウェブ 2 件目	8.2	6.5%
さらにウェブを見る	6.8	5.4%
ウェブ 3 件目	6.2	5.0%
ニュース 1 件目	5.5	4.4%
ウィキペディア 2 件目	3.7	3.0%
ニュース 2 件目	3.7	2.9%
ニュース 3 件目	2.9	2.3%
ウィキペディア 3 件目	2.6	2.0%
さらにウィキペディアを見る	2.1	1.7%
さらにニュースを見る	1.8	1.5%
期間 2(12/11-2/29，81 日間) の平均)		
ウェブ 1 件目	1066.0	30.4%
ニュース 1 件目	464.4	13.2%
ウェブ 2 件目	357.7	10.2%
さらにニュースを見る	326.1	9.3%
ニュース 2 件目	283.7	8.1%
ウェブ 3 件目	253.0	7.2%
ウィキペディア 1 件目	232.4	6.6%
ニュース 3 件目	214.2	6.1%
さらにウェブを見る	136.5	3.9%
ウィキペディア 2 件目	71.4	2.0%
ウィキペディア 3 件目	46.5	1.3%
つながルート	**45.4**	1.3%
さらにウィキペディアを見る	10.0	0.3%

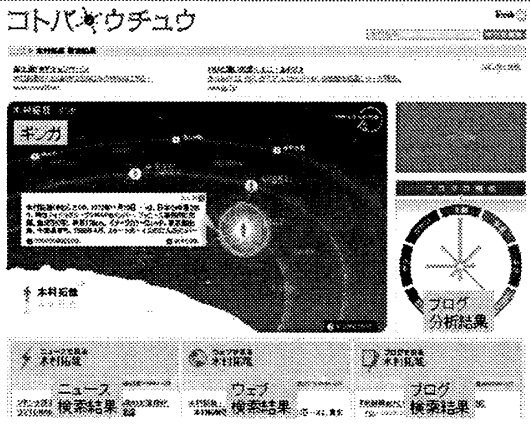

図 4: ギンガと検索結果の画面

てもユーザが各検索結果内のランクを重んずるという事実は興味深い．

4. 探検型検索サイト「コトバ ノ ウチュウ」

前章で述べたように，プロトタイプのクエリログおよびクリックスルーを分析した結果，プロトタイプでは Yahoo!ニュースからの流入クエリの大半に対して「つながルート」が提示されないため，ユーザの情報要求の変化促進ができていないことがわかった．この知見を活かし，探検型検索サイト「コトバ ノ ウチュウ」では，以下の手段により「つながルート」に相当するインタフェース (図 4 のような宇宙に浮かぶ星のメタファを用いた「ギンガ」) のクエリに対するカバレージを広げることにした．

1. 人名以外の主要な Wikipedia エントリも参照関係視覚化の対象とする．現在，約 50 万件の日本語 Wikipedia エントリのうち，約 47 万件をカバーしている．

2. クエリと Wikipedia エントリ名がほぼ完全一致した場合だけギンガを提示するのではなく，部分一致

表 2: Yahoo! JAPAN ニュース検索からの高頻度クエリに対するギンガのカバレージ

(a) プロトタイプで既にカバーできていた (人名のみ)	54	11.2%
(b) 完全一致でギンガが出力されるようになった	130	27.0%
(c) 部分一致でギンガが出力されるようになった	54	11.2%
(d) ニュース共起に基づくギンガが出力されるようになった	22	4.6%
(e) ワンクリックすれば「注目のギンガ」が出力される	222	46.0%
高頻度クエリ異なり数	482	100%

にも対応する．例えばクエリ「エレベータ」に対して,「エレベータ」を中心ノード,「東芝エレベータ」などを周辺ノードとしてもつギンガを提示する．

3. 部分一致にも失敗した場合，Wikipedia ではなく最新のニュース記事を利用し，クエリタームと文内共起した語を周辺ノードとしてもつギンガの提示を試みる．この場合，クエリタームに相当する Wikipedia エントリは存在せず，周辺ノードに相当する Wikipedia エントリのみが存在することになる．

4. 以上に失敗した場合，ユーザにさらにワンクリックさせた上で，クエリとは直接関係はないが話題性の高い「注目のギンガ」を提示する．ここで，話題性の高い語は，ニュース・ブログ・テレビ書き起こしテキストを時系列に並べた上で，比較的古いデータを不適合文書，新しいデータを適合文書と見なして適合性フィードバックを行い選出する．

表 2 に，プロトタイプ検索サイトで収集した前述の異なり数約 10 万のクエリログの中から，頻度が 100 以上のクエリを抽出し，これら全 482 件に対する現状のギンガのカバレージを評価した結果を示す．表中の (a) はプロトタイプで既に「つながルート」が表示できていたものであり，全て人名クエリである．(b) は「コトバ ノ ウチュウ」で新たにギンガが表示可能となったもので，組織名・作品名・イベント名などが多い．(c) は，前述の部分一致によりギンガが表示可能となったものである．(d) は，クエリと Wikipedia エントリの完全一致および部分一致には失敗したが，最新のニュース記事からクエリと文内共起する Wikipedia エントリ名を抽出できた場合である．これは現状では Wikipedia 参照関係抽出の補助的機能という位置づけであるが，今後カバレージを拡大したい．また，(e) は 2 語以上のクエリを含み，現状の仕様ではギンガを提示できていないが,「注目のギンガ」の提示により現在のクエリとは直接関係のない新しい情報要求を誘発する試みである．

表 2 からかわるように，高頻度クエリに対するプロトタイプのカバレージが 11.2%であったのに対し,「コトバ ノ ウチュウ」のカバレージは大幅に向上している．完全一致によるギンガのカバレージ ((a)+(b)) は 38.2%であり，部分一致およびニュース共起によるギンガ出力まで含めると 54.0%である．今後，日本語 Wikipedia 全体を「コトバ ノ ウチュウ」に取り込み，ニュース共起など Wikipedia 参照関係以外の情報をさらに活用すれば，ユーザの興味を惹く周辺星を提供できる確率はさらに高まる可能性がある．

5. まとめ

本研究は，Wikipedia の参照関係を視覚化したクリッカブルなインタフェースを検索結果画面内に提示することにより，ユーザの情報要求の変化を誘発させる探検型検索サイト構築に関する取り組みの第一報である.「コトバ ノ ウチュウ」開発に先立ち，プロトタイプを Web 上で公開し，クエリログとクリックスルーを分析することにより，以下の知見が得られた．

1. Yahoo!ニュースから流入した高頻度クエリのうち，Wikipedia の人名エントリと一致するものの割合はわずか 2.7% (= 2844/107263) 程度であり，多くのユーザに対して Wikipedia の参照関係視覚化により情報要求の変化を誘発させるには，視覚化の対象を人名以外の Wikipedia エントリに拡張することが必須である．

2. ニュース・Web・Wikipedia 各検索結果を同一画面上に配置した検索結果インタフェースにおいても，ユーザは各検索結果中の上位のページをクリックする傾向がある．

上記第 1 の知見をもとに，2008 年 3 月 25 日より Web 上で試験公開している「コトバ ノ ウチュウ」は，Yahoo! ニュースから流入した高頻度クエリの半分以上についてギンガを提示することが可能となっている．今後は「コトバ ノ ウチュウ」自体のクエリログとクリックスルーを用いた評価およびユーザビリティ評価を行い，検索結果提示方法や検索有効性を含めたシステム改善を行う予定である．特に，ユーザの情報要求の変化促進という，従来の検索エンジンにない側面については重点的に評価・改良を行いたい．

参考文献

[1] Bates, M. J.: The Design of Browsing and Berrypicking Techniques for the Online Search Interface, *Online Review*, 13(5), pp. 407-424 (1989).

[2] 想− IMAGINE Book Search: `http://imagine.bookmap.info/index.jsp`

[3] Joachims, T. *et al.*: Evaluating the Accuracy of Implicit Feedback from Clicks and Query Reformulations in Web Search, `ACM TOIS` 25(2), Article No. 7 (2007).

[4] 中山, 原, 西尾: 自然言語処理とリンク構造解析を利用した Wikipedia からの Web オントロジ自動構築に関する一手法, DEWS 2008 (2008).

[5] 連想検索エンジン reflexa: `http://labs.preferred.jp/reflexa/`

[6] Taylor, R. S.: Question-Negotiation and Information Seeking in Libraries, *College and Research Libraries*, 29(3), pp. 178–194 (1968).

[7] Völkel, M. *et al.*: Semantic Wikipedia, *ACM WWW 2006 Proceedings* (2006).

[8] White, R. W. *et al.*: Supporting Exploratory Search, *Communications of the ACM*, Vol. 49, No. 4, (2006).

[9] White, R. W. *et al.* (eds.): Proceedings of the ACM SIGCHI 2007 Workshop on "Exploratory Search and HCI: Designing and Evaluating Interfaces to Support Exploratory Search Interaction" (2007).

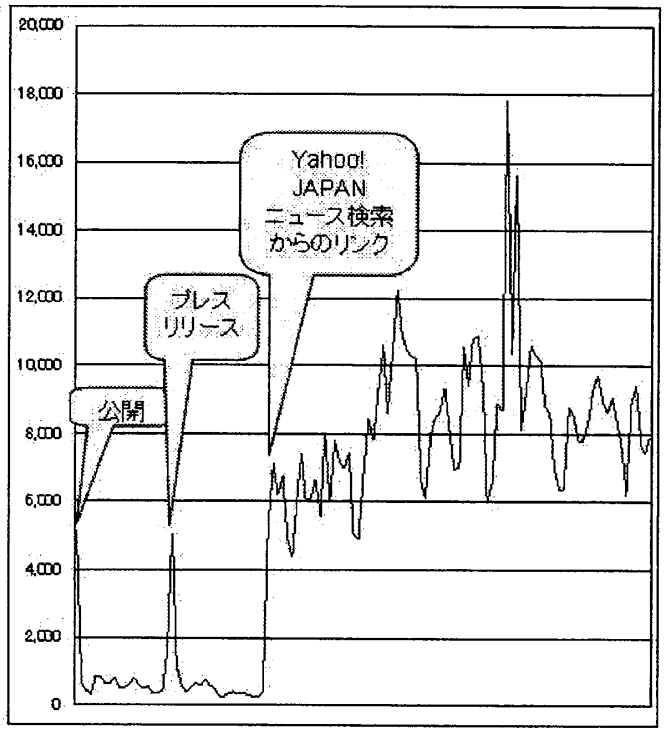

図 2: プロトタイプ検索の日別ページビュー (11〜2 月)

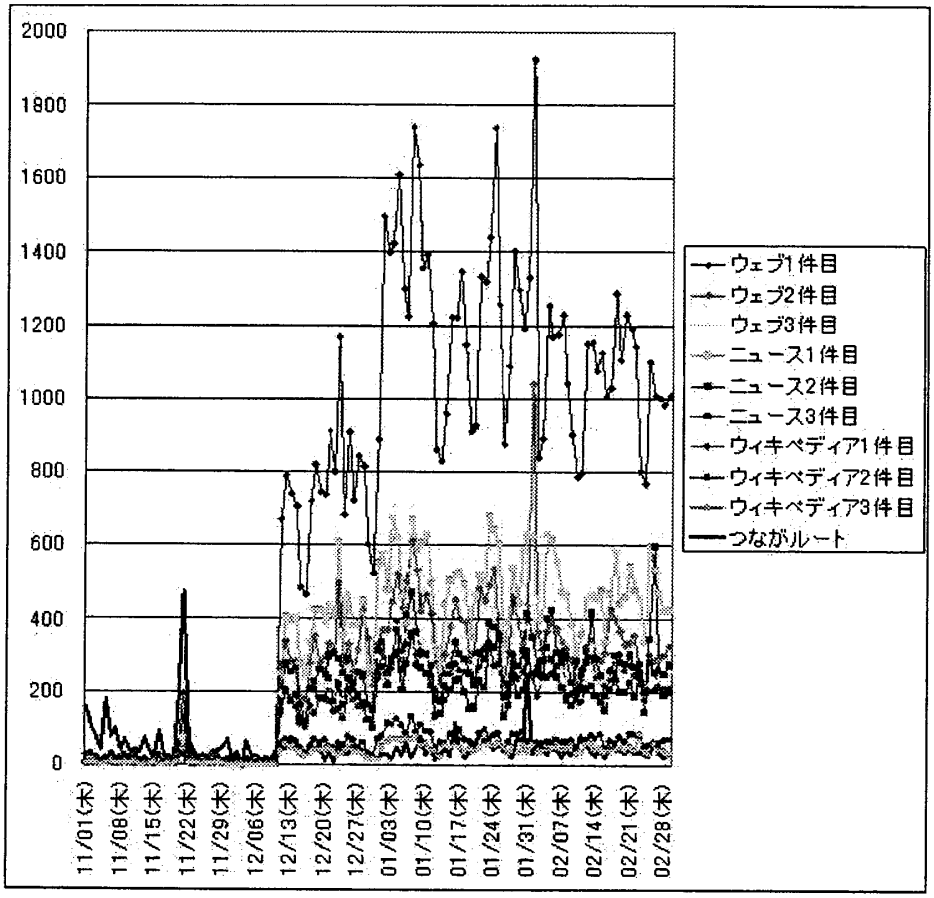

図 3: プロトタイプ検索の日別クリックカウント (11〜2 月)