Rerank-By-Example：編集操作に基づく検索結果の網羅的閲覧

Rerank-By-Example: Browsing Web Search Results Exhaustively Based on Edit Operations

山本　岳洋♥　中村　聡史♦
田中　克己♦

Takehiro YAMAMOTO　Satoshi NAKAMURA
Katsumi TANAKA

通常，検索エンジンが返す検索結果は膨大な数であり，ユーザはそのうちの上位数件しか見ないことが多い．しかし検索タスクの中には情報を網羅的に収集しなければならないような状況も存在する．このような場合，ユーザは自分にとって明らかに不必要な検索結果を幾度もチェックする必要がある．我々はこれまでに，ユーザの編集操作に基づいて検索結果をリランキングするシステムを提案してきた．本稿ではそうした検索結果閲覧タスクに対して本研究の手法を適用し，その有用性を検証する．

Search engines return a huge number of Web search results, and the user usually checks merely the top 5 or 10 results. However, users sometimes have to collect information exhaustively. In this case, the user must check search results repeatedly that are clearly irrelevant to the user. We have proposed a reranking system based on the user's edit-and-propagate operations. In this paper, we apply this idea to that information retrieval tasks and evaluate its usefulness.

1. はじめに

近年各種の検索エンジンを利用した情報収集が一般化している．通常の情報検索では，ユーザは自分にとって必要な結果が数件程度集まれば良いことが多く，検索結果を上位5～10件程度閲覧すれば十分な量の情報を得られることが多い．しかし，情報検索タスクの中には網羅的な検索を行わなければならない状況も存在する．例えば，新しい車を購入するために，ウェブ検索を用いて情報を集めようとする場合には，様々な車種について，性能や評判といった情報をできるだけ多く収集する必要があるだろう．また，自分や他人の書いた論文をすべて集めてくるような際には，論文情報検索エンジンにアクセスし，膨大な検索結果を取捨選択する必要がある．

通常の網羅的検索タスクは大きく分けて下記に挙げる二つの行為が存在し，互いに行き来しながらユーザは求める情報を集めていると考えられる．

検索結果の網羅的収集： 求める情報をできるだけ多く取得することができるように，複数のクエリを作成し，検索エンジンから検索結果を網羅的に取得する行為

検索結果の網羅的閲覧： 検索エンジンから取得した検索結果を網羅的に閲覧し，求める情報を取捨選択する行為

収集した検索結果の網羅的閲覧について考えると，現状では以下の点に問題がある．

検索結果の精度： 情報検索におけるユーザの検索意図は様々なものがある．検索エンジンがユーザの入力した検索クエリだけを利用してユーザの検索意図を把握することは困難であり，検索結果の多くがユーザの求めるものと異なることは珍しくない．そのため，ユーザは網羅的に情報を集める際，明らかに不適合な検索結果を何度もチェックしていかなければならない．

検索結果の分類： 網羅的検索では，検索結果集合から得られた情報を分類することが重要である．最も単純な分類は，この検索結果は自分が求める情報である，または求める情報ではないという分類である．それ以外にも，例えば，論文のサーベイをしている場合に，自分の研究と関係がありそうな論文の検索結果を，分野や研究の関連度で分類するといったことが挙げられる．ユーザはこのような分類を行う際，頭の中ですべて記憶しておくとか，関係のある検索結果のページのウィンドウを開いたままにしておくであるとか，メモをとる等の手間をかけなければならない．

上記の問題点から，現状の網羅的検索タスクでは，検索結果の適合・不適合の判断，検索結果からリンク先ページへの行き来，検索結果の分類，クエリの生成といった種々の操作が入り乱れ，ユーザの負担となっている．

我々はこれまでに，削除や強調といった編集操作を可能とし，検索結果をリランキングするシステムを提案してきた[1]．本稿では，これまでに我々が開発してきたシステムをさらに拡張し，網羅的検索の閲覧タスクを支援する方法を提案する．提案するシステムを利用することで，ユーザは検索結果の精度改善と検索結果の分類を，編集操作を用いて容易に行うことができるようになる．

2. 網羅的検索タスク

2.1 網羅的検索タスクとは

網羅的検索とは，調べたい事柄に関連する集合的な知識を，１つないし複数の観点からできるだけ漏れが無いように収集していくようなタスクである．そのようなタスクでは，検索結果の収集と，その検索結果の網羅的閲覧という行為が欠かせない行為であると考えられる．

2.2 網羅的収集

ユーザは網羅的検索タスクにおいて，求める答えを含むようなクエリを生成しなければならない．

このタスクを支援する手法としては，クエリ拡張[2]のような手法を用いて，ユーザの求める答えをできるだけ含むようなクエリを自動的に生成することが有効であると考えられる．本システムでは，ユーザは手動でクエリを入力するが，クエリ拡張のような手法は網羅的収集を支援する上で有効であると考えられる．

2.3 網羅的閲覧

検索結果を網羅的に閲覧する上での問題点は，1 章で述べたとおり，検索精度の改善と，検索結果の分類という２種類

♥ 学生会員　京都大学大学院情報学研究科　博士前期課程
tyamamot@dl.kuis.kyoto-u.ac.jp
♦ 正会員　京都大学大学院情報学研究科社会情報学専攻
{nakamura, tanaka}@dl.kuis.kyoto-u.ac.jp

に分けることができる．大規模な検索結果を効果的に分類する手法として代表的なものにはクラスタリング[3][4]が挙げられる．[5]にあるように，文献データベース等のクラスタリングは，非常に高い精度で同姓同名人物を判定でき，ある人物の論文情報をすべて集めたいといった網羅的検索タスクに対して非常に有効に働くことも考えられる．しかし，そのクラスタ内の結果を，自分の研究の関連度や論文の分野で分類する際，クラスタリングの手法だけでは困難である．ユーザのニーズに合った検索結果の精度向上や，ユーザの求める評価軸での検索結果の分類のためにはユーザとシステムのインタラクションが不可欠である．また，インタラクションによりクラスタリングの性能も飛躍的に向上させることができると考えられる．

3. 編集操作を用いた網羅的閲覧支援

3.1 インタラクション

ユーザとシステム間のインタラクションは図１の通り．

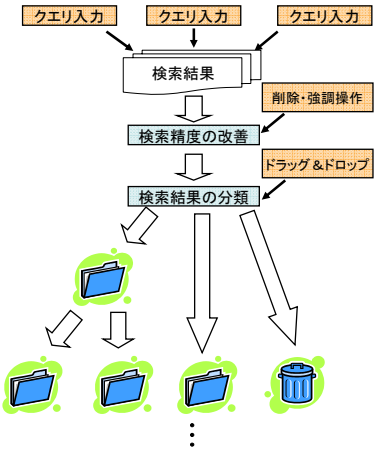

図１：ユーザとシステム間のインタラクション
Fig.1 Interactions between users and our system

（１） ユーザはシステムにクエリを入力
（２） システムはクエリを検索エンジンに送り検索結果を取得
（３） ユーザは収集した検索結果を順次閲覧
（４） 必要に応じて以下の３種類の操作を再帰的に行う
　(a) 編集操作を用いた検索精度の改善
　(b) 編集操作を用いた検索結果の分類
　(c) クエリの追加による新たな検索結果の収集
（５）ユーザは検索結果を十分に分類し終わると，分類された検索結果や未分類の検索結果を閲覧し，求める答えを獲得する

3.2 検索精度の改善

3.2.1 操作の種類

ユーザの行う編集操作の種類は以下の通りである．

削除操作： 削除操作は，ユーザが自分にとって不要な情報をシステムに伝える操作である．システムは削除されたキーワードや検索結果と関係のある検索結果を下位にリランキングする．

強調操作： 強調操作はユーザが自分にとって必要な情報をシステムに伝える操作である．システムは強調されたキーワードや検索結果と関係のある検索結果を上位にリランキングする．

3.2.2 操作の対象

一般に，検索結果はタイトルやURL，スニペットといった属性から構成される．また，論文の検索結果ではこれらに加えて，著者リスト，学会名，掲載年度といった属性も検索結果を構成する要素となる．

本論文では，このような属性中のキーワードに対する操作による検索結果のリランキングを実装した．

3.2.3 リランキングの流れ

ユーザのキーワードの削除・強調に基づいた検索結果のリランキングの流れを以下に示す．

（１）システムはユーザに検索結果集合$R=\{r_1, r_2, \ldots, r_N\}$を提示（$r_i$は$i$位の検索結果）
（２）ユーザは受け取った検索結果を順次閲覧し，必要に応じて検索結果中の属性$attr$中のキーワードsを削除・強調
（３）システムはユーザが編集操作を行った属性$attr$とキーワードsを取得し，属性$attr$内にsを検索結果内に含むような検索結果 $r \in R$ に対して以下の式を適用する

$$SC(r) = i + type * N \quad (1)$$

ここで，$SC(r)$は検索結果rの適合度，$type$ は操作の種類であり，削除の場合は-1，強調の場合は1となる．またN は検索結果の数を表す．

（４）検索結果を$SC(r)$の値でリランキングしユーザに提示
（５）（２）へ戻る

3.3 検索結果の分類

一般的なウィンドウシステムでは，ファイルを分類・整理するためにフォルダという機能が一般に用いられている．複数のファイルを一つのフォルダに格納することで，複数のファイルをまとめて扱うことができる．そこで，本稿では，検索結果を分類・整理するための機能として，このフォルダを導入する．

3.3.1 操作の種類

ユーザはドラッグ＆ドロップという様々なアプリケーションの中で広く利用されている編集操作を用いて検索結果の分類を行う．ドラッグ＆ドロップは，ある場所にあるものを，より適切な場所へと移動する操作である．例えば，不要なファイルをゴミ箱へ捨てるであるとか，文書の編集中にコンテンツを別の適切な場所へと移動するといったことに利用している．本稿では，ドラッグ＆ドロップを，ユーザが分類したい検索結果をシステムに伝える操作として導入する．

3.3.2 操作の対象と検索意図

キーワード： ユーザが検索結果中のあるキーワードをあるフォルダに入れるということは，ユーザはそのキーワードを含むような検索結果をそのフォルダへと分類したいという意図を汲み取ることができる．

検索結果集合： 自分の求める検索結果集合を上位にリランキングした後，適合する結果を全て，あるフォルダに分類したいということがある．その場合，ユーザは該当する検索結果をすべて選択し，フォルダへとドラッグ＆ドロップすると考えられる．

フォルダ： フォルダはユーザが分類した検索結果の集合である．複数のフォルダの検索結果を併合したいときは，ユーザはあるフォルダを別のフォルダへとドラッグ＆ドロップすると考えられる．

4. 実装

提案システムを実現するため，システムの実装を行った．本システムはウェブ検索結果と論文情報検索結果を扱うことができる．ウェブ検索エンジンとしては Google，論文情報検索エンジンとしては GoogleScholar を利用した．

システムの実行例を図 2 に示す．システムはクエリ入力エリア，検索結果表示エリア，及び検索結果分類エリアから成り立っている．

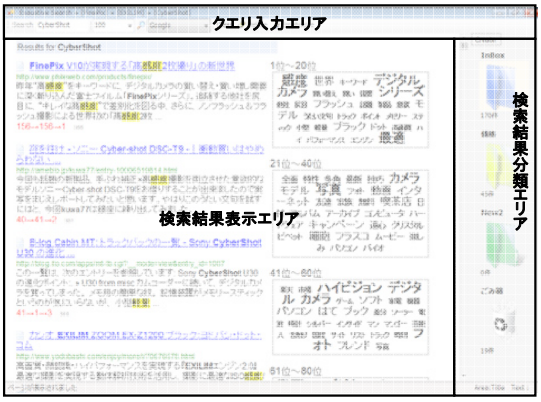

図 2 ：システムの実装例
Fig. 2 Screenshot of our system

4.1 クエリ入力エリア

ユーザはクエリ入力エリアにクエリを入力して検索結果を収集する．ユーザがクエリを入力すると，システムはクエリを検索エンジンに送り，検索結果を取得する．システムはその検索結果を検索結果分類エリア最上段のフォルダに追加し，検索結果を表示する．この仕組みにより，ユーザは一つのタスクで複数のクエリの検索結果を統一的に扱うことが可能となる．

4.2 検索結果表示エリア

検索結果表示エリアでは，ユーザは編集操作を用いて検索結果のリランキングを行うことができる．ユーザが検索結果中のキーワードを選択すると，削除と強調を行う操作ボタンがマウスカーソル周辺に現れる．同時に，ユーザが選択したキーワードを含む検索結果がハイライト表示される．これはユーザが選択したキーワードを含む検索結果がどの程度存在するのかを，視覚的に表すものである．その後ユーザが押したボタンに基づきシステムは検索結果のリランキングを行う．

また，検索結果表示エリアの右側部分には，検索結果集合の傾向の提示を行うためタグクラウドを表示している[1]．ユーザはこのタグクラウドに対しても編集操作を行うことができる．

4.3 検索結果分類エリア

検索結果分類エリアはいくつかのフォルダから成り立っている．初期状態として，ユーザの入力したクエリの検索結果をはじめに格納するフォルダ（エリア内上部）と不適合な検索結果を入れるごみ箱フォルダ（エリア内下部）が用意されている．ユーザが検索を開始するか，ユーザが明示的にフォルダの追加を行うと新しいフォルダが生成される．ユーザは新しく生成されたフォルダにラベル付けを行うことができ，分類を容易にすることができる．なお，ラベル付けが行われていないフォルダにキーワードや複数の検索結果がドラッグ＆ドロップされると，システムはそのフォルダ内の検索結果内に頻出するキーワードをいくつか選び，自動的にフォルダのラベル付けを行う．

ユーザはキーワードや検索結果を，ドラッグ＆ドロップを用いてフォルダに入れることにより検索結果の分類を行う．また，フォルダをクリックすることで，フォルダに分類されている検索結果を閲覧することができる．

4.4 実行例

論文情報検索エンジンから，“京都大学の教授である田中克己氏が著者として入っている論文”を集めたいというタスクを例に，システムの実行例を説明する．図 3 にはこのタスクにおけるシステムの図を示した．まず，ユーザはクエリとして「“K Tanaka” Web 」と入力し，検索結果を受け取る．

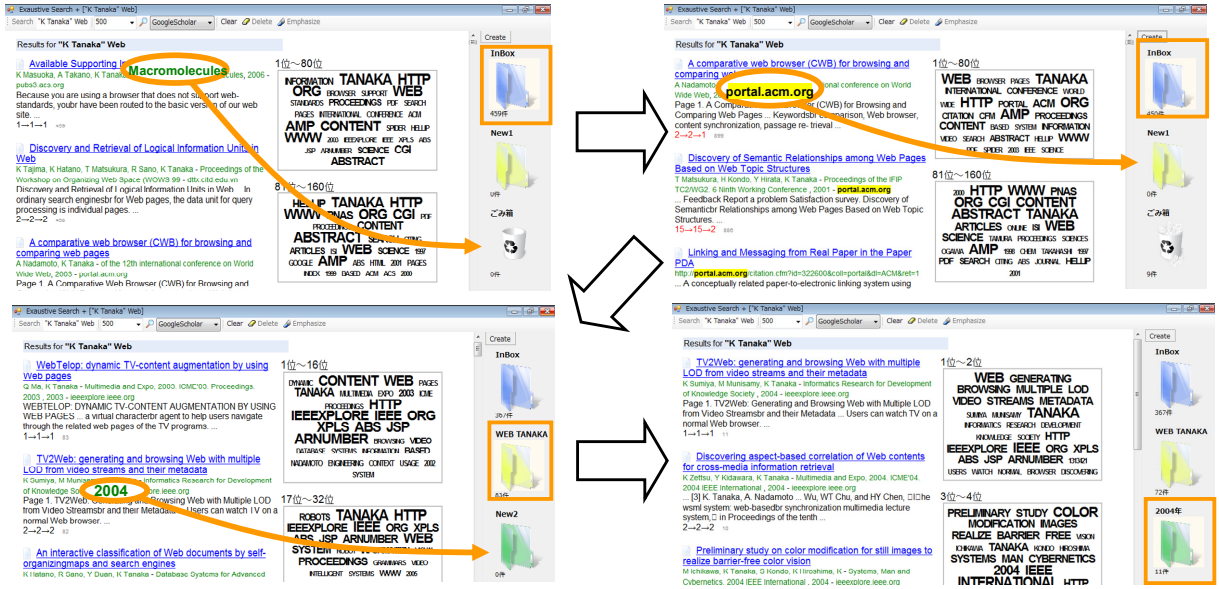

図 3 ：システムの実行例
Fig. 3 Example of our system

そして，検索結果を閲覧していき，関係のなさそうな語（ここでは Macromolecules という会議名)を選択し，ごみ箱フォルダへとドラッグ＆ドロップする．次に，関係のありそうな語（ここでは portal.acm.org，ieeexplore.ieee.org，www.springerlink.com という論文掲載サイト）を強調し，リランキングされた検索結果を確認する．上位にリランキングされた結果のほとんどが関係ありそうであれば，これらの語を選択し，フォルダにドラッグ＆ドロップする．次に，分類を行ったフォルダ内の検索結果を表示し，例えば，ユーザが 2001 年に発表した論文を見てみたい思えば，2001 という語を選択し，別のフォルダへと追加すると，新しいフォルダには 2001 年の，しかも自分が求める人物と関係のある論文の検索結果が高い精度で分類される．

5. 実験

本システムを使用した検索結果集合の分類の効率性を調べるため，実際にシステムを用いて検索結果を分類する実験を行った．タスクの流れを以下に示す．

まず，ユーザは与えられたクエリの検索結果 500 件をシステムから受け取る．その後，ユーザは検索結果集合を，あらかじめ決められた分類に従って分類する．実験に用いたタスクは，正解セットの作成の容易さから，同姓同名人物の分類や多義語の分類を行い．今回は 1 つのクエリに対し 2 つの分類を用意した．なお，分類は全てキーワードのドラッグ＆ドロップによって行い，分類された検索結果集合の再現率が 90%を超えるか，操作回数が 8 回を超えた時点でユーザはその分類を終了する．このタスクを 6 個のクエリについて行い，操作回数の増加による適合率と再現率の推移を調べた．

なお，再現率は，検索エンジンで得られる上位 500 件内にすべての答えが含まれていると仮定し，算出した．

図 4 は 6 つのクエリで分類された 12 種類の検索結果集合の適合率と再現率の平均である．12 個の分類のうち，8 回の操作では再現率が 9 割まで達しなかった分類が 3 つ存在した．また，3 回以下の操作で再現率が 9 割を超えたものは 3 種類あった．図から分かるように，3 回～4 回の操作を行えば，再現率が 8 割を超えることが分かる．

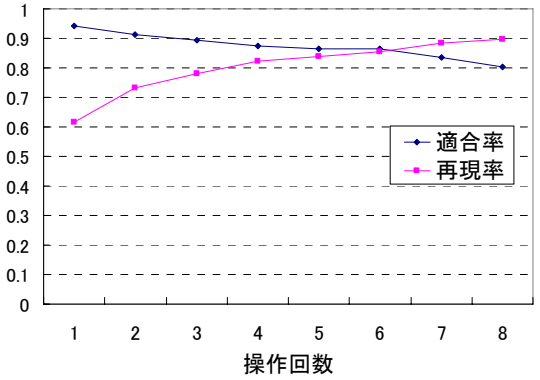

図 4：操作回数と適合率・再現率の推移
Fig.4 Precision and recall of each operation

6. まとめ

本稿では，編集操作を用いて，網羅的検索を支援するシステムを提案した．本稿では網羅的検索タスクを網羅的収集及び網羅的閲覧という二つの行為に分類し，特に網羅的閲覧を支援する手法を提案した．本システムでは削除・強調操作を行うことによって検索結果をリランキングし，ドラッグ＆ドロップを行うことによって検索結果を分類する．今後はどのようなタスクにおいて本システムが効果的に働くのかをユーザ実験によって検証する必要がある．

我々は今後，オンラインショッピングのような，機械的に生成されるウェブサイドについても，今回提案したリランキングや分類を行うことが可能なシステムについて検討していく予定である．

[謝辞]

本研究の一部は，文部科学省研究委託事業「知的資産の電子的な保存・活用を支援するソフトウェア技術基盤の構築」，異メディア・アーカイブの横断的検索・統合ソフトウェア開発（研究代表者：田中克己），文部科学省科学研究費補助金特定領域研究「情報爆発時代に向けた新しい IT 基盤技術の研究」における，計画研究「情報爆発時代に対応するコンテンツ融合と操作環境融合に関する研究」（研究代表者：田中克己，A01-00-02，課題番号 18049041），計画研究「情報爆発に対応する新 IT 基盤研究支援プラットフォームの構築」（研究代表者：安達淳，Y00-01，課題番号：18049073），文部科学省科学研究費補助金若手研究(B)No.18700129，および，文部科学省グローバル COE 拠点形成プログラム「知識循環社会のための情報学教育研究拠点」（研究代表者：田中克己，平成 19～23 年度）によるものです．ここに記して謝意を表すものとします．

[文献]

[1] T. Yamamoto, S. Nakamura and K. Tanaka: “Rerank-By-Example: Efficient Browsing of Web Search Results”, Proc of DEXA2007, to appear.
[2] J. Xu and W. Croft: “Query expansion using local and global document analysis”, Proc of SIGIR1996 , pp. 4–11 (1996).
[3] D. Hand, H. Mannila and P. Smyth: “Principles of Data Mining”, Bradford Book (2001).
[4] H. Zeng, Q. He, Z. Chen, W. Ma and J. Ma: “Learning to cluster web search results”, Proc of SIGIR2007, pp. 210–217 (2004).
[5] M. Bilenko, R. Mooney, W. Cohen, P. Ravikumar and S. Fienberg: “Adaptive name matching in information integration”. Intelligent Systems, IEEE, 18, 5, pp. 16–23 (2003).

山本 岳洋 Takehiro YAMAMOTO
京都大学大学院情報学研究科修士課程在学中．日本データベース学会学生会員．

中村 聡史 Satoshi NAKAMURA
京都大学大学院情報学研究科社会情報学専攻特任助教．2004 年大阪大学大学院情報学研究科博士後期課程修了．博士（工学）．主にヒューマンコンピュータインタラクション，ウェブ検索の研究に従事．情報処理学会，日本データベース学会会員．

田中 克己 Katsumi TANAKA
京都大学大学院情報学研究科社会情報学専攻教授．1976 年京都大学大学院修士課程修了．博士（工学）．主にデータベース，マルチメディアコンテンツ処理の研究に従事．IEEE Computer Society，ACM，人工知能学会，日本ソフトウェア科学会，情報処理学会，日本データベース学会等各会員．