C-15

検索ランキングを考慮した Web 検索結果の網羅的閲覧

Exhaustive Browsing of Web Search Results with Ranking

厚見 悠太†　　大島 裕明†　　田中 克己†
Yuta Atsumi　　Hiroaki Ohshima　　Katsumi Tanaka

1. はじめに

近年のインターネットの発展はめざましく，情報を取得する際には Web 検索エンジンが用いられることが非常に多くなっている．ユーザが明確な検索意図をもって検索を行う場合には，適合する文書が取得できれば良い．そのように，明確に適合する文書が決定できる場合には，望む Web ページを得るために，適合フィードバック[1]などの手法が利用可能である．しかし，ある話題の情報について網羅的に情報を得たい informational search の場合には，現在の Web 検索エンジンのランキングに従った検索結果のリスティングだけでは不十分である．

例えば，ユーザが就職活動を行うために，就職活動に関連する様々な事柄について調べたいと考えている場合を想定する．ユーザが得たい情報は，面接におけるマナー，就職活動に向けて準備すべきこと，他の人の就職活動の様子など，非常に多岐にわたる情報である．このような多様な情報が 1 つのページで網羅されていることはほとんどないため，たいていの場合は多くの異なるページを閲覧する必要が生じる．しかし，現在の Web 検索エンジンが返す検索結果は，主にクエリに対する適合度とページの重要度に基づくランキングを行っており，多様な内容をなるべくもれなく調べるために，検索結果の上位から 1 件 1 件すべてのページを調べていくことは想定していないと考えられる。例えば，検索結果内には重複した内容についてのページが多く含まれており，網羅的に調べる際には，重複する内容を選別しながら閲覧する必要があり，そのことはユーザには大きな負荷となってしまっている．

本論文では，あるクエリによって得られる検索結果のページ集合から，できるだけ少ない閲覧量で，多様な情報を取得する手法について提案する．この目的を達成するには様々な手法が考えられる．我々は，既存の検索エンジンが出力した検索結果を上位から順番に閲覧することを想定し，上位のページの内容から，あるページを閲覧する必要があるかないかを，そのページの内容の既閲覧ページに対する新規性（novelty）と網羅性（coverage）を用いて判定する手法を提案する．提案手法は，より上位に順位付けされたページに類似のものが存在したり，すでに内容が包含されると考えられたりする場合には，対象のページは閲覧しなくても良いと判定するものである．検索エンジンはクエリとの適合度やページの重要度などから検索結果の各ページに対して順位を与えている．その順位を維持したまま検索結果を閲覧することにはユーザが日常的に行っていることであるが，本論文で提案する手法は，その順位を維持したまま読まなくて良いページが判断できるため，ユーザは通常の検索エンジンの利用から違和感なく網羅的閲覧が可能になると考えられる．

† 京都大学大学院情報学研究科社会情報学専攻
〒 606-8501 京都府京都市左京区吉田本町

2. 関連研究

検索結果を網羅的に閲覧することに関連する研究やサービスは，いくつか存在している．

湯本ら[2]はユーザの求める情報の全容を表すページ集合を発見する全容検索を提案している．全容を表すページ集合は結果として網羅性が高いページ集合である可能性が高いので，本研究との関連は深い．しかし，この研究では，あくまでクエリの典型的なトピックをなるべく多く網羅することを主眼に置いており，結果として出力されるのは数ページからなるページ集合である．本研究では，検索結果全体の内容を網羅することを目的としており，目的が異なると考えられる．

Zhang ら[3]は検索結果のランキングにおいて，文書の多様性と文書の持つ情報の豊かさという 2 つの指標を考慮する手法を提案した．提案された手法でランキングされた検索結果の上位を閲覧することで，なるべく多様な情報が得られるという点で本研究と関連がある．類似の文書を除去する手法に関連する研究としては Zhang ら[4]は新規性と冗長性を考慮した文書検索について研究を行っている．与えられたクエリとは関連（relevance）があることに加えて，すでに閲覧した文書など，他の文書が網羅する内容に対して冗長性（redundancy）がなく，かつ，新規性（novelty）がある文書を発見する手法を提案している．これらの手法は，文書のランキング手法の一種である．我々の手法は，既存の検索エンジンがあたえた順位を維持したまま，網羅性を向上させる閲覧を実現することを目的としており，ランキング手法の提案ではない．文書検索以外にも，画像検索では Song ら[5]が多様性について考慮したランキング手法を提案している．

Matsuike ら[6]は検索結果のページ集合の概要を内包的にキーワード集合で表現する手法を提案している．この研究では概要をキーワード式と呼ばれるキーワードを論理式で表した式によって表現している．提案されているキーワード式は，検索結果全体において，どのような内容が存在するかを表すと考えることもできる．そのような技術は網羅性を評価する上で必要になってくると考えられる．

山本ら[7][8]は検索結果を語ベースの適合フィードバックによって，検索結果を再ランキングする手法を提案している．これは，ユーザが検索結果を閲覧しながら，必要な語や不要な語を指定すると，動的に検索結果が並び変わるというインタラクティブなシステムである．我々の手法は静的な検索結果を出力するものであり，異なるアプローチであるが，検索結果の網羅的閲覧を実現する手法としては関連するものである．

3. 検索結果の網羅的閲覧

本節では，検索結果の網羅的閲覧とはどのような問題であるかを提起し，本研究はどのような位置付けであるかを述べる．

3.1 検索結果の網羅的閲覧における問題

この問題の入力は，検索エンジンが出力した検索結果である．検索結果はランキングされたページの集合であるが，より正確には，検索結果に含まれるページの集合$P = \{p_1, p_2, \ldots, p_n\}$とその集合における全順序関係$\leq$の組からなる全順序集合$(P, \leq)$と表現できる．それに対する出力としては多様な形態が考えられるが，例えば，ページ列，ページを要素とするクラスタ，文書の要約などが考えられる．

この問題に必要な要件は 2 つある．1 つは入力として与えられた検索結果から多様な情報を得ることであり，もう 1 つは利用者が閲覧する量をできるだけ抑えることである．この 2 つはトレードオフの関係にあるので，両方を実現することは一般に困難である．

そこで，評価軸として閲覧量と網羅量という 2 つの軸を設定する．閲覧量はユーザが閲覧したページや文書の量であり，網羅量は検索結果全てに含まれる情報の内，どれだけ網羅できたかを示す量である．

この問題は閲覧量を低くしつつ網羅量をできるだけ高くすることが最終目的となる．図 1 に本研究の目的を表す図を記す．この図の意味は、左側は従来の検索エンジンの閲覧量と網羅量の関係を表しており、右側は本研究が目的とするシステムの閲覧量と網羅量の関係を表している。

3.2 ページ列の出力問題

前項で定義した問題のうち，出力がページ列である場合について考える．出力がページ列の場合，問題はさらに以下の 3 つに分類することができる．

1. Pの要素である全ページの再ランキング問題
2. Pの部分集合に対する再ランキング問題
3. Pの部分順序集合を出力する問題

1 つ目は，P全体を出力する場合である．この場合，出力のページ列は入力の検索結果における順序$\leq$を保つことはない．なぜなら，順序$\leq$を保った場合，入力の検索結果そのものが出力されることになるためである．そのため，この問題はP全体における新たな全順序関係を求めること，すなわち，Pの要素であるページの再ランキングを行うことになる．文献[3][4][5]が対象としているのはこの問題である．

2 つ目は，1 つ目と同様にPの要素に対する再ランキングを行うが，出力されるのはPの部分集合である場合である．我々は，以前この問題に取り組んだ[9].

3 つ目は，$(P, \leq)$の部分順序集合，すなわち，Pの部分集合を出力し，順序として$\leq$を保つ場合である．本論文で取り組む問題が，これに該当する．

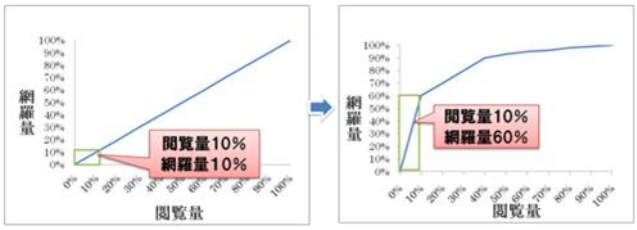

図 1 検索結果の閲覧量と内容の網羅量の関係性

このような，ページ列を出力とする場合，評価における閲覧量は閲覧したページ数となり，網羅量は検索結果に含まれる全てのトピックの内，出力結果が含んでいるトピック数と定義することができる．よって，順番に閲覧していくと多様なトピックを含むページが閲覧できるようなページ列求めることが，問題に対する目的となる．

3.3 入力における順序を保持したページ列の出力問題

本研究では，前項で示したページ列の出力問題のうち，入力の検索結果のランキングを保持する場合を考える．すなわち，入力の検索結果のランキングである順序関係$\leq$を前提として，入力のページ列を$P = (p_1, p_2, \ldots, p_n)$と表現すると、出力のページ列$R$は$P$の部分列となる．

検索結果のランキングを保持することが有効に働く例はいくつか考えられる．例えば，ユーザが従来の検索エンジンで返される検索結果のページを閲覧し，上位 10 件を閲覧したとする．そこで，残りのページを網羅的に閲覧したいと考え，提案手法によって，それ以降のページを閲覧すべきかどうかを判定した場合を想定する．この場合，従来の検索エンジンで返される結果と提案手法によって返される結果は順序が変動していないので，ユーザに混乱が生じないという利点がある．また，ユーザが閲覧した 10 件より下位のランクのページは上位のランクのページを考慮して不要ではないページであることが保障されているので，ユーザがまだ得ていない情報が得られる確率が高い．

4. 提案手法

本節では，本論文における提案手法が基づく仮定と，手法の詳細について述べる．

4.1 本研究の手法における仮定

4.1.1 類似性に関する仮定

本研究の目的の 1 つは情報を多様にすることである．そのため，一度閲覧した内容と同じ内容について述べた文書を再び閲覧することはなるべく避ける必要がある．我々の以前の研究[9]では，ユーザが既に読んだページを入力として与え，それに基づいて重複したページを除去していた．

本研究では検索結果は上位から閲覧していくものであるので，上位のページと類似しているページはユーザが一度閲覧した既読のページであると見なすことができる．よってそのようなページを閲覧する必要がないと考え，削除の対象とする．

4.1.2　包含性に関する仮定

ページ間の包含性について考える．例をあげると，「就職活動」というクエリで検索した場合，面接時のマナーだけ書いてあるページと，就職活動全般におけるマナーが書いてあるページでは包含関係が成り立つ．後者のページの方がここでは重要なページであると見なされる．

一般に，あるページに書かれてある話題が，他のページに書かれてある話題に含まれている時，後者のページは前者のページを包含していると定義し，このページ間に包含関係があると定義する．

本研究では個々のページ同士の包含関係を 1 つ 1 つ考えるのではなく，多くの包含関係を持つページを重要なページと見なす．そのために，多くの文書に現れる単語を含むページは，多くの文書を包含する可能性が高く，重要性が高いという仮定を置く．この仮定により，網羅性の低いページは削除の対象となり，多くのページを包含するページのみが出力の対象となる．

4.2　出力のページ列の生成手順

検索エンジンで返されるランキングされたページ列を$P=(p_1,p_2,\ldots,p_n)$と表現する．ただし，nは対象とする検索結果ページの件数である．Pが入力として与えられた時，出力となるランキングされたページ列Rを以下のように作成する．

1. Rをp_1のみを要素とするページ列とする．すなわちR=(p_1)となる．
2. k = 2とする．
3. p_kに対してScore(R,p_k)を計算し，その値が閾値θ以上であればp_kをRの最後の要素として追加する．
4. kの値を 1 増加させ，kがnになるまで手順 3 を繰り返す．

このランキングされたページ列作成の手順を図 2 に記す．この図は，k = 4の場合を表している．すでに，p_1とp_3はページ列Rの要素となっている．p_2は内容的にp_1に含まれるものであったため，スコアが閾値を上回らず，ページ列Rの要素とはなっていない．この状態では，スコアはp_4とページ列Rの要素であるp_1とp_3から計算される．Score(R,p_k)の計算方法は次節で述べる．

4.3　スコアの計算

Score(R,p_k)を以下の式で定義する．

$$\mathrm{Score}(R,p_k)=\alpha\mathrm{Nov}(R,p_k)+(1-\alpha)\mathrm{Cov}(p_k) \quad (1)$$
$$(0\le\alpha\le 1)$$

この式はページ p_k の新規性を表すNov(R,p_k)と網羅性を表すCov(p_k)を足し合わせたものである．

αは新規性と網羅性のどちらを重視するかを決める定数であり，0 から 1 の間の値を取る．

Nov(R,p_k)は以下の式で定義される．

$$\mathrm{Nov}(R,p_k)=1-\max_{r_i\in R}(\cos(r_i,p_k)) \quad (2)$$

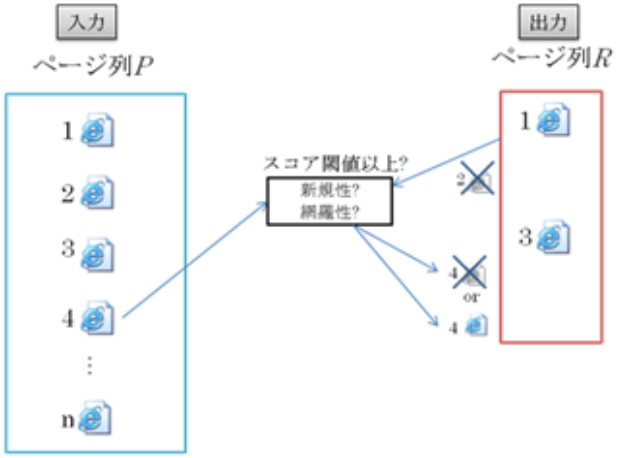

図 2　出力のページ列の生成手順

Nov(R,p_k)はページの新規性を表す尺度である．Rには定義上，既にユーザが読んだページが含まれているので，それぞれのページとの類似度をはかってその最大値を 1 から引いたものがNov(R,p_k)となる．この値は大きいほど，対象とするページ p_kは新規性が高いページとなる．この尺度は 0 から 1 の間の値を取る．

Cov(p_k)は以下の式で定義される．

$$\mathrm{Cov}(p_k)=\frac{\mathrm{DF}(p_k)}{\max_{2\le i\le n}\mathrm{DF}(p_i)} \quad (3)$$

Cov(p_k)は網羅性の尺度である．対象とするページ $\mathrm{p_k}$ が含んでいる単語が，どれくらい他のページに現れる可能性が高いかを計算している．多くのページに含まれる単語を多く含んでいるページほど，網羅性が高いと言えるので，Cov(p_k)が高いほど，網羅性が高いページと言える．網羅性が高いページほど，多くのページを包含する可能性が高く，結果として多くのページを包含するページが出力されることとなる．この尺度は正規化されているので，0 から 1 の間の値を取ることとなる．

cos はコサイン類似度を表す．ページp_i,p_jに含まれる単語を特徴量とする特徴ベクトルをそれぞれv_i,v_jとする時，コサイン類似度は以下の式で定義される．

$$\cos(p_i,p_j)=\frac{v_i\cdot v_j}{\|v_i\|\|v_j\|} \quad (2)$$

DF(p_k)は，DocumentFrequency のことであり，以下の式で定義される．

$$\mathrm{DF}(p_k)=\sum_{t_i\in p_k}\mathrm{DF}(t_k) \quad (3)$$

DF(t_k)は単語$\mathrm{t_k}$が含まれる文書の数で定義される．

Nov(R,p_k)も Cov(p_k)も 0 から 1 の値をとるので，Score(R,p_k)は 0 から 1 の間の値をとることとなる．

5. 終わりに

本稿では検索結果の網羅的閲覧の問題設定を行い，その要件である順序付きページ集合の提示手法を提案した．

今後の課題として挙げられるのは，従来の検索エンジンの検索結果や，関連研究の手法を用いて提示されるページ集合と本手法で提示されるページ集合を比べる必要があるということである．システムを実装して，テストセットを作成し，客観的な評価を行いたい．

謝辞

本研究の一部は，京都大学グローバル COE プログラム「知識循環社会のための情報学教育研究拠点」，および，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」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，課題番号 18049041）によるものです．ここに記して謝意を表します．

文献

[1] J. Rocchio: Relevance feedback in information retrieval, In the SMART retrieval system: experiments in automatic document processing, Prentice-Hall (1971).

[2] T. Yumoto, K. Tanaka: Page sets as Web search answers, Proc. of the 9th International Conference on Asian Digital Libraries, pp.244-253 (2006).

[3] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W. Fan, Z. Chen, W-Y. Ma: Improving Web search results using affinity graph, Proc. of the 28th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.504-511 (2005).

[4] Y. Zhang, J. Callan, T. Minka: Novelty and redundancy detection in adaptive filtering, Proc. of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 81-88 (2002).

[5] K. Song, Y. Tian, T. Huang, Wen Gao: Diversifying the image retrieval results, Proc. of the 14th Annual ACM International Conference on Multimedia, pp.707-710 (2006).

[6] Y. Matsuike, S. Oyama, K. Tanaka: Proc. of the 6th International Conference on Web Information Systems Engineering, pp.607-608 (2005).

[7] 山本岳洋，中村聡史，田中克己: Rerank-By-Example: 編集操作に基づく検索結果の網羅的閲覧，日本データベース学会論文誌（DBSJ Letters），Vol.6, No.2, pp.57-60 (2007).

[8] T. Yamamoto, S. Nakamura, K. Tanaka: Reranking and classifying search results exhaustively based on edit-and-propagate operations, Proc. of the 20th International Conference on Database and Expert Systems Applications (2009) (to appear).

[9] 厚見悠太，大島裕明，田中克己: Web ページの既読結果を用いた Web 検索結果の網羅的閲覧，第 1 回データ工学と情報マネジメントに関するフォーラム予稿集, A3-1 (2009).