クエリに関するトピック量とユーザの総閲覧量を考慮したWeb検索ランキング

厚見　悠太†　　大島　裕明†　　田中　克己†
京都大学大学院　情報学研究科社会情報学専攻†

1.はじめに

近年，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ち，取得したい情報を限定的に意識して検索を行う場合には，適合するページが取得できれば良い．しかし，ある話題の情報について網羅的に情報を得たい場合には，ユーザが漠然とWeb検索エンジンが返す結果を上位から逐一読んでいく，などといったことを行う必要があり，ユーザにとって負荷が高いと考えられる．

そこで，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を再ランキングし，その部分集合を出力することで，ユーザに負荷をかけることなく検索結果に含まれるトピックを網羅する手法について提案する．

我々は，ユーザが少数のページを望む場合は広く浅い網羅が適切になり，多数のページを望む場合では深い網羅が適切になることに注目し，ユーザに望む総閲覧量を事前に入力させてどちらの種類の網羅を望んでいるかを判断する．

提案手法は，対象とするページに含まれるトピック数を用いて，ページのスコアを計算した後，ページに含まれるトピックの重複をさけるために，高いスコアを持つページとの類似性を考慮してスコアを再計算している．このスコアを用いてページ集合を再ランキングしてユーザが入力した件数だけ出力する．

2.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．

2.1 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$P=\{p_1,p_2,...,p_n\}$である．それに対する出力としてはページ列やクラスタなど多様な形態が考えられる．

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できるだけ多くのトピックを網羅することであり，もう1つは利用者が閲覧する量を抑えることである．

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．同じ閲覧量であれば，網羅量が高いほど良い評価となる．

2.2 本研究の位置付け

本研究では出力をページ列とし，Pを再ランキングした後に部分集合を出力する問題に取り組む．この問題の再ランキングに関する部分はB. Zhangら[1]やY. Zhangら[2]も取り組んでいるが，本研究ではユーザの望む網羅の形態について考え，代表度と網羅度という観点を導入している所に相違点がある．

2.3 閲覧量，網羅量の定義

本論文での閲覧量はユーザが指定したページ数となる．網羅量はユーザが望むページの性質を考慮することで，以下のようになる．

(1) 全ページ数$|P|$における各トピックの代表ページの合計数$|P_{rep}|$の内，閲覧したページに含まれる代表ページ数$|p_{rep}|$の割合

(2) ページに含まれる全てのトピック数$|T|$の内，閲覧したページに含まれるトピック数$|t|$の割合

1つ目の定義では$|p_{rep}|/|P_{rep}|$で網羅量を計算し，2つ目の定義では$|t|/|T|$で網羅量を計算する．

これはユーザが検索エンジンで返される検索結果を網羅したい場合に，各トピックの深い知識を欲する場合と，広く浅い知識を欲する場合があることに起因している．

3.提案手法

本節では，本論文における手法の詳細について述べる．

3.1.解析手法

Web Search Ranking based on the number of query's topics and a user's browsing pages
†Department of Social Informatics, Graduate School of Informatics, Kyoto University

入力として与えられた検索結果をページ列 $P=\{p_1,p_2,...,p_n\}$ とする．各ページの形態素解析を行い，単語を次元として重みを *tf - idf* 値とする特徴ベクトルをそれぞれ $V=\{v(p_1),v(p_2),...,v(p_n)\}$ とする．この特徴ベクトルを用いて各ページのスコアを計算する．

3.2 ページのスコアリング

$Score(p_k)$ を以下の式で与える．

$$Score(p_k)=\begin{cases}Rep(p_k)\{\text{if } input > totaltopic(P)\}\\ Cov(p_k)\{\text{if } input < totaltopic(P)\}\end{cases} \quad (1)$$

ただし *input* はユーザが入力した総閲覧量を表わし，$avgtopic(P)$ は以下の式で計算する．

$$totaltopic(P)=\sum_{p_i\in P}(topic(p_i)) \quad (2)$$

この式は入力した総閲覧量がトピック数の合計によって，スコアが代表度に基づいたものになるか網羅度に基づいたものになるか変化することを表わしている．

代表度は，対象とするページがどれだけ各トピックに関して深い内容を記述しているかを表わす値である．本論文では，暫定的な代表度を以下の式とする．

$$Rep^n_{temp}(p_k)=\frac{1}{|topic(p_k)|}*\frac{length(p_k)}{\max_{p_i\in P}(length(p_i))} \quad (3)$$

ただし $length(p_k)$ はページ p_k に含まれる単語数とする．また，$|topic(p_k)|$ はページ p_k に含まれるトピック数を表わしている．詳細は後述する．この式では，1ページに含まれるトピック数が少なく，多くの単語が含まれているページほど値が高くなる．

網羅度は，対象とするページが全体に対してトピックをどれだけ網羅しているかを表わす値である．本論文では，ページの暫定的な網羅度を以下の式とする．

$$Cov^n_{temp}(p_k)=|topic(p_k)| \quad (4)$$

この式では，1ページに含まれるトピック数が多いページほど値が高くなる．

3.3 トピック数の導出方法

本論文ではトピック数を以下の式で計算する．

$$|topic(p_k)|=|\{t_i \mid v(p_k)[t_i]>\theta\}| \quad (5)$$

この式では，ページ p_k に含まれる単語の内，*tf - idf* 値が閾値を超えるものをトピックと見なしている．

3.4 代表度，網羅度の再計算

代表度，網羅度の再計算を行う．これは重複するトピックを持つページのスコアを下げるために行う．

p_k よりスコアが高いページ集合を p_{upper} とすると，再計算は以下の式で再帰的に行われる．

$$Rep^{n+1}_{temp}(p_k)=(1-\max_{p_i\in P_{upper}}(sim(p_k,p_i)))Rep^n_{temp}(p_k) \quad (6)$$

$$Cov^{n+1}_{temp}(p_k)=(1-\max_{p_i\in P_{upper}}(sim(p_k,p_i)))Cov^n_{temp}(p_k) \quad (7)$$

ここでの $sim(p_k,p_i)$ はページの類似度を表わしており，本論文では以下の式で計算する．

$$sim(p_i,p_j)=\cos(v(p_i),v(p_j))=\frac{v(p_i)\cdot v(p_j)}{\|v(p_i)\|\times\|v(p_j)\|} \quad (8)$$

i 回の再計算の後，$Rep(p_k)$ と $Cov(p_k)$ を決定する．

3.5 出力方法

前項までで計算された $Score(p_k)$ に基づいてページをランキングする．そして上位のページからユーザが入力した件数だけ出力する．

4.終わりに

本論文では検索結果の網羅的閲覧という問題の中で，ユーザの総閲覧量を考慮しながら，ページの代表度，網羅度に基づいたスコアで再ランキングをするという手法を提案した．今後は評価実験を行い，提案手法の妥当性を示して行きたい．

謝辞

本研究の一部は，京都大学 GCOE 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，A01-00-02，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] B. Zhang, H. Li, Y. Liu, L. Ji, W. Xi, W.Fan,Z. Chen and W. Ma: "Improving web search results using affinity graph", Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp.504–511 (2005).

[2]Y. Zhang, J. Callan and T. Minka: "Novelty and redundancy detection in adaptive filtering", ACM Press, pp. 81–88 (2002).