Web ページ集合を解とするランキング手法

A Ranking Method Returning Page Sets as Web Search Answers

湯本 高行♥ 田中 克己♦

Takayuki YUMOTO Katsumi TANAKA

検索技術の向上により，ユーザの求める情報が以前より容易に見つけられるようになっているが，それが当てはまらない場合がある．例えば，ユーザの欲する情報が複数の場所に分散して存在する場合である．それに対して我々はページ集合を検索の解とする統合型ランキングを提案する．さらに，ランキングの対象となるページセットを効率よく生成する手法について説明する．また，統合型ランキングを全容検索に応用する手法について述べる．全容検索とは与えられたキーワードに対する全容を記したページ(集合)を発見することを目的とした検索である．本稿ではグラフを用いてページセットの内容やページ間の関係を表現する手法を提案する．さらに統合型ランキングと他の既存のランキング手法を比較する．

Progress of search technology made it easier for users to get information which they want but it is sometimes still difficult. For example, it is the case where information which they want is distributed on multiple Web pages. To solve this problem, we propose page set ranking to evaluate page sets as ranking units. We explain algorithm to generate pertinent page sets efficiently. We also apply the page set ranking to overview search. Overview search is to find pages which describe about overview of given keywords. We propose graph-based model to express content of page set and relationship between pages. We compare our ranking method and conventional ranking methods.

1. はじめに

近年，インターネットの普及や検索技術の向上により，ユーザの求める情報が以前より容易に見つけられるようになっているが，それが当てはまらない場合がある．例えば，ユーザの欲する情報が複数の場所に分散して存在する場合である．このような場合，現在の検索エンジンはページ毎に判断しているため，ランキングの上位に似たようなページが固まってしまい，効率よく情報を発見することは難しい．これに対して我々はページ集合を検索の解とする統合型ランキングを提案している[1,2]．統合型ランキングでは与えられたキーワードを用い，既存の検索エンジンで検索を行い，取得したページを組み合わせて適するページセットを生成し，それらをランキングする．本稿ではそのアルゴリズムを説明する．さらに，与えられたキーワードに対する全容を記したページ(集合)を発見する全容検索を提案し，統合型ランキングをそれに応用する手法について述べる．

♥ 学生会員 京都大学大学院情報学研究科博士後期課程
yumoto@dl.kuis.kyoto-u.ac.jp
♦ 正会員 京都大学大学院情報学研究科
ktanaka@i.kyoto-u.ac.jp

2. 統合型ランキング

2.1 概要

統合型ランキングはページだけではなく，それらを組み合わせたページセットについてもランキングの対象とする検索手法である．統合型ランキングでは，入力は検索結果ページのランクつきリスト(現在の手法ではランクなしの集合でもよい)であり，出力はページセットのランクつきのリストである．統合型ランキングの目的は本稿で述べるような全容検索だけではなく，ある2つの事柄を異なる視点から比較したページを集める比較検索[1]などさまざまであるが，その目的に応じてランキング計算に用いる評価値や制約が異なる．統合型ランキングでは既存の検索エンジンを用いてページを取得し，それらを組み合わせてページセットを生成，ランキングする．各ページセットは，ページセットが表現する内容に関する特徴量とページセットがどのようなページから構成されるかについての特徴量を持ち，これらを用いてランキングの評価値やページセットが満たすべき制約は表現される．図1に統合型ランキングのイメージを示す．

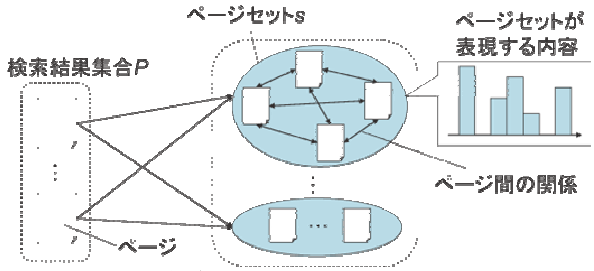

図1 統合型ランキングのイメージ
Fig.1 Image of Page Set Ranking

2.2 ランキングの計算

任意のページの組み合わせをランキングの対象とした場合，計算量が$O(2^n)$かかってしまうため，ランキングの対象とするページセットの数を削減する必要がある．そこで，ランキングの対象を極大なページセットのみに限定する．極大なページセットとはランキングの対象が満たすべき制約条件を満たすようにページセットにページを追加していったときに，これ以上ページを追加すると，制約条件を満たさなくなるページセットである．また，ランキングの対象のページセットが満たすべき制約条件とは別に，ページセットの生成時にランキングの評価値の増加が一定の水準(閾値θ_{diff})以上であるという条件も付加し，ランキングの対象の数を制限する．以上をまとめると次のようなアルゴリズムになる．

1. $R \leftarrow \phi$, $S_1=\{\{p_1\},\{p_2\},\cdots,\{p_n\}\}$, i=1とする．
2. $S_{i+1} \leftarrow \phi$とする．
3. $s \in S_i$に対して，$s'=s \cup \{p\}$を計算する．
   (ただし, $p \in P, p \notin s$)
   $score(s')-score(s) \geqq \theta_{diff}$かつ制約条件を満たす場合
   $S_{i+1} \leftarrow S_{i+1} \cup \{s'\}$とし，それ以外は$R \leftarrow R \cup \{s\}$とする．
4. $S_{i+1} \neq \phi$ならば，$i \leftarrow i+1$として，2に戻る．
5. $s \in R$について，評価値に基づきランキングを計算する．

ただし, {p}は単一のページpからなるページセット, s∪{p}はページセットsにpを追加してできるページセット, score(s)はページセットsの評価値である. また, p∈sはページセットsがページpを含むことを意味する. 上記の制約条件や評価値は目的によって異なり, ページセットが表現する内容に関する特徴量とページセットがどのようなページから構成されるかについての特徴量を用いて表現される.

2.3 関連研究

Clusty[3]などクラスタリングを利用した検索エンジンが提案されている. クラスタリング結果の構造はユーザがほしい情報を効率的に選ぶ際にある程度の助けになるが, ユーザに予備知識が全くない場合, 有効に使うことはできない. また, 統合型ランキングでは異なる情報を含むページからページセットを生成する必要があるが, クラスタリングによってページの組み合わせの数を削減できる可能性がある. しかし, 各クラスタからどのページを選択すればよいかをクラスタリングの技術だけで知ることはできないので, クラスタリングの技術だけでは分散する情報を効率よく収集することは難しい.

また, 田島らはリンクでつながったWebページやネットニュースやメールのスレッドに対して, 検索の解として適切な範囲を抽出する手法を提案している[4]. この手法はリンクなどによって既に関連づけられているものに対して解の粒度を大きくして適した解を発見しようというものである. これに対して, 提案手法は粒度を大きくするだけではなく, 関連づけられていない任意のコンテンツを組み合わせて適した解を構成するものであり, より適した解を発見できる可能性がある.

3. 全容検索

3.1 概要

全容検索は与えられたキーワードについて全容がわかるようなページ(集合)を発見することを目的とする. 話題の広さと深さの両方を兼ね備えているときに全容を表現しているものとする. 全容検索は, 例えば全く知識がない分野についてのサーベイを行う場合などに有効である. もし, 全容を1ページで記述したページが存在すれば, 統合型ランキングを用いる必要はないが, 複数のページから情報を集めることで全容がわかることも少なくない. そこで, 統合型ランキングの評価値を後述する被覆度とし, 与えられたキーワードの全容をどの程度カバーしているかを表現する. さらに, 閲覧の効率を考えるとページ間の内容の重複はなるべく避けるべきであるので, 制約条件として後述する重複度が一定の水準(閾値 θ_{dup})以下という条件を設定する.

3.2 詳細グラフ

検索結果集合 P における語の出現状況から語の関係を表す詳細グラフを作成する. この詳細グラフを用いて, ページセットの特徴量を定義し, ランキングに用いる.

3.2.1 詳細関係

語によって, 概要的な内容を記述したページに出現しやすい, 詳細を記述したものに出現しやすいなどの傾向があり, これらの違いを表現するために, 検索結果集合での語の出現状況から語の詳細関係を定義する. 検索結果集合にあまり出現しない語はそのトピックにおいてあまり重要でないと考え, 検索結果集合 P に含まれる語のうち, DF の上位 m 位に含まれる語のみを詳細語の候補とし, 詳細語候補集合 T とする.

共起度を $cooc(t_i,t_j)=DF(t_i,t_j)/DF(t_i)$ と定義する. ただし, $DF(t_i, t_j)$は語 t_i, t_j を共に含む文書数, $DF(t_i)$は語 t_i を含む文書数である. ここで, 語 t_j が語 t_i より詳細である($t_i \prec t_j$)状態は, 語 t_i, t_j が同時に出現する文書数が十分ある中で, 語 t_j が含まれる文書中で, 語 t_i が出現する確率が高く,その逆は言えない状態と定義する. さらに, 詳細関係は推移すると定義する. この条件を式で表すと以下のようになる.

$DF(t_i, t_j)/|P| > \theta_{DF}$

$cooc(t_j,t_i) > \theta_{cooc} \wedge cooc(t_i,t_j) < \theta_{cooc}$

$\forall t \prec t_j, t \prec t_i$ … (A)

式(A)より詳細関係 $\prec$ は同じページ集合 P において推移性が成り立つ半順序関係である.

3.2.2 グラフの定義

この詳細関係を利用して, 上位に概要的な語が位置し, 枝を辿るにつれ, より詳細な語に至るようなグラフ, 詳細グラフを定義する. 各語をノードとし, 以下が成り立つ場合に t_i から t_j へ有向枝を張るものとする.

$t_i \prec t_j \wedge \nexists t, t_i \prec t, t \prec t_j$

クエリ q によって得られた検索結果集合内では, どのページにも必ずクエリ q は含まれるので, このグラフのルートはクエリ q をキーワードに持つ.

さらに以下が成り立つとき, t_i と t_j は密接な関係にあると考えられるので, ノードを集約し, t_i,t_j に対応するノードの親子関係を引き継ぐ.

$cooc(t_j,t_i) > \theta_{cooc} \wedge cooc(t_i,t_j) > \theta_{cooc}$

また, 一般的な語 t_c が候補語集合に紛れ込んでいる場合, その語が詳細グラフの上位のノードと認識されるおそれがある. そのため, 小山らによる詳細語判定のアルゴリズム[5]により, t_c が詳細語の候補かどうかを判定し, 詳細語でなければ除外する. 問い合わせ回数削減のため, 本稿では, ルートに直接つながっているノードが持つキーワードについてのみ詳細語かどうかを判定する. 例えば, 図 2 でブダペストというラベルのついたノードはルートに直接つながっているが, ドナウというラベルがついたノードは直接つながっていない.

このようにしてクエリ"ハンガリー"に対して図 2 のような DAG が得られる.

3.3 被覆度

詳細グラフにおいて, 共起関係から上位のノードに含まれるキーワードは内容を概要的に表す語, 下位のノードは内容を詳細に表す語と考えることができる. 下位のノードに含まれるキーワードについて詳細語か否かを検証しないのは, そのキーワードが一般的な語であった場合, 出現数が多いと考えられるので, IDF による重みづけを行うことにより, 影響を少なくできるからである.

c(s,t)はページセット s に語 t が含まれているかを表す関数で, ページセット s が語 t を含むとき 1, それ以外は 0 である. 重みつき被覆度 cov_w を以下のように与える.

$$cov_w(s, g_q) = \frac{1}{|child(n_q)|} \sum_{n \in child(n_q)} cov_{node}(s, n)$$

$$cov_{node}(s, n) = \frac{\sum_{t \in g_n} IDF(t) c(s,t)}{\sum_{t \in T} IDF(t)}$$

g_q はクエリ q によって得られた詳細グラフである. n_q はクエリ q をキーワードに持つノードであるが, 詳細グラフは検索

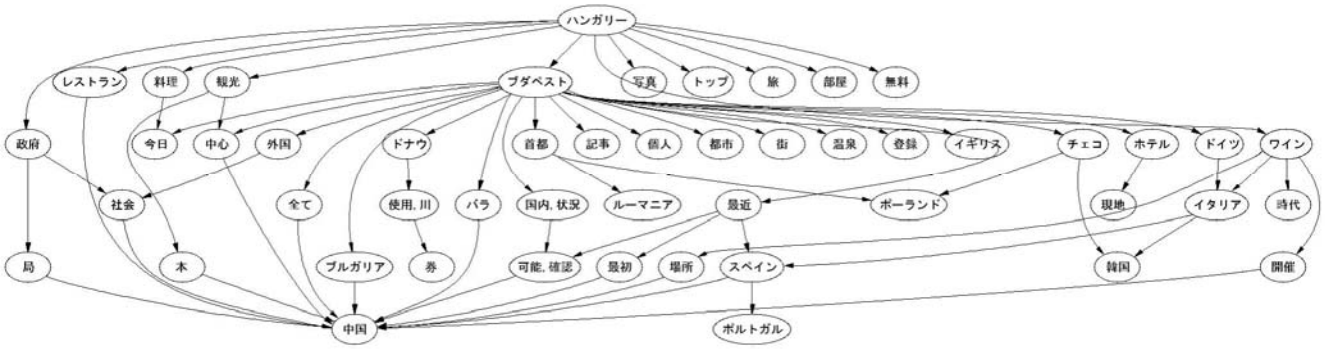

図 2 詳細グラフの例
Fig.2 Example of Detailing Graph

結果集合から生成することを前提するため，すべてのページはクエリとして与えたキーワード q を含むので，ルートは必ず q をキーワードに含む．つまり，n_q はここではルートである．child(n_q)はノード n_q に直接つながっている子ノードの集合とする．

IDF は IDF(t)=log(|P|/DF(t))+1 と定義する．ただし，DF(t)は P において，語 t を含む文書数，|P|は P に含まれるページの数である．

ノード n∈child(n_q)をルートとする各部分グラフはクエリ q の主要なサブトピックを表していると考えられる．cov_{node} は，ページセットが各部分グラフに含まれる語のどのくらいの語を含んでいるか，つまり対応するサブトピックの内容をどの程度含んでいるかを[0,1]で表している．語の重みとして IDF を採用し，DF の小さいもの，つまり詳細な内容を説明するために用いられている語の重みを重くしている．これによって全容検索の目的である，内容の広さと深さの両方を兼ね備えたページセットを表現している．

3.4 重複度

重みつき重複度 dup_w を以下のように定義する．

$$dup_w(s,g_q)=\frac{1}{|child(n_q)|}\sum_{n\in child(n_q)} dup_{node}(s,n)$$

$$dup_{node}(s,n)=\frac{\sum_{t\in g_n} IDF(t)ov(s,t)}{\sum_{t\in T} IDF(t)}$$

ただし，g_n は詳細グラフ g のノード n をルートとする部分グラフ，ov(s,t)は，s 内の複数のページに語 t が含まれているときには 1，それ以外のときには 0 を返す関数である．

重みつき重複度についても重みつき被覆度と同様に，各サブトピックにおいて，複数のページに含まれる語がどのくらいあるかを表している．dup_{node} では，DF の小さいもの，つまり詳細グラフ内の下位に近いノードの重みを重くしている．これは DF の大きいものは概要について述べていると考えられるので，内容が異なる文書に出現する頻度が高いと考えられ，重みは低くするべきと考えられる．また，DF の小さいものは詳細について述べており，内容の異なる文書には出現する頻度が低く，このような語が重複して出現する場合は同じ内容を述べていると考えられるので，重みは高くするべきという考えに基づいている．また，dup_w でもそれぞれのサブトピックを同じ重みで扱っている．

3.5 アルゴリズム

全容検索の処理の流れは大きく分けて以下の 3 段階である．

- 検索結果集合の取得
- 詳細グラフの計算
- ランキングの計算

検索結果集合の取得では，既存のページ毎の検索エンジンにより，与えられたキーワードについて検索を行い，ランキングの上位 m 件を取得し，P とする．

ランキングの計算においては，制約は $dup_w(s')<\theta_{dup}$，評価値は $cov_w(s,g_q)$ とし，2.2 で述べたアルゴリズムによってランキングの計算を行う．

4. 実験

4.1 実験環境

全容検索の検索エンジンには，Google[6]を用い，検索結果の上位 100 件を取得し，検索結果集合 P とし， その中での DF 値の上位 100 語を詳細語候補集合 T とした．パラメータは，θ_{cooc}=0.8，θ_{dup}=0.5，θ_{DF}=0.2 とした．また,ページセットを構成するページ数は |s|≦3 に限定して実験を行った．

4.2 評価

全容検索の全容検索の評価を行うため，(1)全容検索によって検索したページセットのランキング，(2)検索結果を重みつき被覆度でリランキングし，上位 3 件をページセットとみなしたもの，(3)Google の検索結果の上位 3 件をページセットとみなしたものを比較する．これらのページセットは(1)提案した手法によって閲覧を行った場合，(2)被覆度の高いページ順に閲覧を行った場合，(3)Google のランキング順に閲覧を行った場合において，ユーザがそれぞれ同じページ数を閲覧した際に，得る情報をページセットとして表現している．4 つのクエリ(鳥インフルエンザ，ハンガリー，風力発電，京都)に対して，比較を行い，平均をとったものが表 1 である．

表 1 統合型ランキングと他の手法の比較
Table 1 Comparison between Page Set Ranking and Other Methods

検索手法	(a)	(b)	(c)
(1)	0.979	0.472	0.415
(2)	0.964	0.866	0.745
(3)	0.463	0.280	0.218

(a)ページセットの重みつき被覆度の平均
(b)ページセットの重みつき重複度の平均
(c)ページの重みつき被覆度の平均

表 2 全容検索の例
Table 2 Example of Overview Search

手法	ページ	タイトル	説明
(1)	Page1-1	鳥インフルエンザについてのQ＆Aを更新しました	一般消費者向け情報
	Page1-2	埼玉県／高病原性鳥インフルエンザに関する情報について	一般消費者/飼育者向け情報
	Page1-3	鳥インフルエンザは大流行するか―バイオ企業の動向	ニュース記事
(2)	Page2-1	鳥インフルエンザ＆新型インフルエンザ情報	リンク集
	Page2-2	宮城県/畜産課/鳥インフルエンザについて	一般消費者/飼育者向け情報
	Page2-3	埼玉県食品安全企画室/鳥インフルエンザに関する対応について	一般消費者/飼育者向け情報
(3)	Page3-1	厚生労働省: 鳥インフルエンザに関する情報 関連情報	トップページ
	Page3-2	国民の皆様へ（鳥インフルエンザについて）	一般消費者向け情報
	Page3-3	国立感染症研究所 感染症情報センター: 鳥インフルエンザ	トップページ

この結果より，提案手法(1)と比較して，(2)では，被覆度は若干劣るものの同等程度の被覆度が期待できるが，重複度が極めて高いため，非効率な閲覧しかできないことが言える．また，(3)では，重複度は極めて低いが，被覆度も低く，必要な情報を得られていないと言える．このような点から提案手法は，内容の被覆度，重複度の両方の面でバランスがとれた検索手法であると言える．

また，各手法でどのようなページが含まれているかを考察するために、表 2 にクエリを"鳥インフルエンザ"としたときの結果のページのタイトルと内容の簡単な説明を記す．ただし, Page1-X は手法(1)によるページセットを構成するページ, Page2-X, Page3-X はそれぞれ手法(2),(3)の場合である．表 2 を補足すると，(1)の場合では，Page1-1,Page1-2 は鳥インフルエンザの FAQ であったが, Page1-1 には主に消費者向けに詳しい説明があり，Page1-2 は消費者向けの情報だけではなく，飼育者向けの情報も含んでおり,お互いに異なる情報を含んでいた．また，Page1-3 は鳥インフルエンザに対する企業の取り組みを紹介しており，他の 2 つとは異なる情報が書かれていた.これに対して，(2)の場合では Page2-1 は鳥インフルエンザについての簡単な説明と数十ページへのリンクが張られていた．また，Page2-2, Page2-3 は共に一般消費者と飼育者に向けた情報が書かれていたが，消費者向けの情報については Page1-1 ほど詳しい内容を含んでいなかった. (3)の場合では，Page3-1,Page3-3 は鳥インフルエンザについてのサイトのトップページでそのページ自体にはあまり情報が含まれておらず，リンク先の個別ページに実際の内容が記述されていた．Page2-1,Page3-1,Page3-3 のようなページでは，リンク先まで閲覧することによって実際の内容が得られると考えられるが，リンク先のページ数がそれぞれ 10 ページ以上と多く，多くのページを閲覧しないと全容の理解は難しく，効率的な閲覧にとっては不利であると言える．

5. おわりに

本稿では，ページセットを検索の解とする統合型ランキングを提案し，ページセットを効率よく生成する手法について説明した．さらに,それを全容検索に応用した．全容検索は，内容的な広さと深さを両立し，ページ間の内容の重複の少ないページ集合を発見する．検索結果集合から詳細グラフを生成して，語の詳細関係を求め，それを用いてページセットを生成するアルゴリズムを示した．また，実験により，従来のページ毎の検索に対する優位性を確認した．今後は，アルゴリズムの高速化，ページセットの表示手法を検討する．

[謝辞]
本研究の一部は，文部科学省 21 世紀 COE 拠点形成プログラム「知識社会基盤構築のための情報学拠点形成」(リーダー：田中克己，平成 14-18 年度) 及び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」，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すものとします．

[文献]
[1] Yumoto, T. and Tanaka, K.: “Page Sets as Web Search Answers”, Proceedings of The 9th International Conference on Asian Digital Libraries (ICADL2006) pp. 244-253 (2006).
[2] Yumoto, T. and Tanaka, K.: “Finding Pertinent Page-Pairs from Web Search Results”, Proceedings of The 8th International Conference on Asian Digital Libraries (ICADL2005), pp.301–310 (2005).
[3] clusty.jp: http://clusty.jp/.
[4] Tajima, K., Mizuuchi, Y., Kitagawa, M. and Tanaka, K.: “Cut as a Querying Unit for WWW, Netnews, and E-mail”, Proc. of ACM Hypertext ’98, pp. 235–244 (1998).
[5] Oyama, S. and Tanaka, K.: “Query Modification by Discovering Topics from Web Page Structures”, Proceedings of the Sixth Asia Pacific Web Conference (APWEB’04), pp. 553–564 (2004).
[6] Google, http://www.google.com

湯本 高行　Takayuki YUMOTO
京都大学大学院情報学研究科博士後期課程在学中．2003 同修士課程修了．主に検索・統合に関する研究に従事．情報処理学会，日本データベース学会， ACM, IEEE 各学生会員．

田中 克己　Katsumi TANAKA
京都大学大学院情報学研究科教授．1976 京都大学大学院修士課程修了．工学博士．主にデータベースの研究に従事．情報処理学会，日本データベース学会，人工知能学会，ACM, IEEE Computer Society 等各会員．