8x-3

# Webページ閲覧のための巡回プラン提示方式の提案

杉山 一成　松本 一則　橋本和夫
株式会社　KDD研究所

## 1　はじめに

WWW上の情報は、急速な勢いで増加している。検索エンジンはユーザが見たいページを探すための一手段であるが、ユーザが入力した検索語に対して多数の候補を出力する。そのため、検索結果から本当にユーザが望んでいるページを見つけたり、ある事柄に対していくつかのページを調べて理解を深めたりすることは、極めて困難になっている。[1]のページ間の参照関係と適合度を用いた検索エンジンとしてgoogle[2]があるが、特定の有名サイトが検索結果のランキングの上位に位置する傾向があることに加え、多数の検索結果が出力されるのは、yahoo[3]やgoo[4]などの従来の検索エンジンと同じである。また、[5]ではユーザの目標概念の理解を目的として、有用なWebページの系列を自動生成するシステムを構築している。しかし、プランの作成に時間を要するため、即座に結果を出すことを要求される検索には不向きであると考えられる。また、ユーザは検索語以外にも、それに関連した文脈語を入力することを要求される。したがって、検索語の意味がわからない場合、ユーザがその語に対する文脈語まで想像するのは困難である。さらに、[7]、[8]はユーザのブラウジングの履歴から、そのユーザが次に見たいと予測される一つのWebページを提示するだけであり、予測が外れた場合にはユーザは検索語に関する内容に関して深い理解を得られないことになる。以上を考慮すれば、ユーザは従来どおり検索語を入力し、システムがWebページのリンク構造に伴う文脈を考慮した上で、ユーザの理解を得られるようにいくつかのWebページの組を出力する形態が望ましいと考えられる。そこで、本論文では、ユーザが入力した検索語に基づき、かつその内容が容易に理解できるようにWebページを閲覧する順番を提示する方式を提案する。

"The proposal of the way of navigation plan for reading Web pages", Kazunari Sugiyama, Kazunori Matsumoto, Kazuo Hashimoto: KDD R&D Labolatories Inc.

## 2　巡回プランアルゴリズムの比較検討

一般的なサイトにおいて、Webページは目次的なページから詳細な内容へと移り変わるという特徴を有している。また、リンクについても参照元のページに関連したものがある一方で、全く内容の違うページへと移るものもある。したがって、こうした特徴を考えれば、Web空間を巡回するためのプランについては、以下の手法が考えられる。

**手法1**：既存のリンクを切り離した上で、巡回プランを提示できるようにリンクの再構成を行う。

**手法2**：検索エンジンの出力結果を用いて、巡回プランを構成する。

**手法3**：既存のリンク構造を活用し、巡回プランを生成する。

手法1は、リンクを切り離しているために、そのリンクによって構成されていた文脈的なつながりが悪くなると考えられる。また、手法2においても、検索エンジンの結果は一度リンクが切り離された断片的なものであり、やはり文脈的なつながりが失われているものと考えられる。一方、手法3では既存のリンクを生かすことで、文脈的なつながりを保つことができ、システムが提示する巡回プランはユーザにとっても容易に理解できるものと考えられる。そこで、次節で述べる巡回プラン生成方式を提案する。

## 3　提案方式

前節で述べたように、巡回プランの生成を既存のリンク構造を活用し、以下の手順で行う。

1. 各Webページの重みベクトルを作成する。ここで、HTMLのタグを利用し、名詞$t$についての重み$w_t$を次式で表す。

$$w_t = \alpha \cdot tf \cdot idf \qquad (1)$$

ただし、タグ情報がない場合には $\alpha = 1$ とする。また、タグ情報がある場合、次のタグで囲まれた名詞について、$\alpha$ は以下の値をとるものとする。

<TITLE> : 10
<Hn> : 6.5-n
<U> : 2
<EM> : 2
<STRONG> : 2

(1) 式を用いて、Web ページ $d_i$ の重みベクトル $\boldsymbol{w}^{d_i}$ は Web ページから抽出された全名詞を $t_1, t_2, \cdots t_n$ とすれば、

$$\boldsymbol{w}^{d_i} = (w_{t1}, w_{t2}, \cdots, w_{tn})$$

と表される。

2. ハイパーリンクが張られている隣接 Web ページ間の類似度を計算し、Web ページ群 $S_c = \{S_{c1}, S_{c2}, \cdots\}$ を構成する。ここで、Web ページ $d_i$ と Web ページ $d_j$ の類似度 $sim(d_i, d_j)$ は、以下のように表される。

$$sim(d_i, d_j) = \frac{\boldsymbol{w}^{d_i} \cdot \boldsymbol{w}^{d_j}}{|\boldsymbol{w}^{d_i}| \cdot |\boldsymbol{w}^{d_j}|}$$

ここで、$v_{th}$ を閾値として $sim(d_i, d_j) > v_{th}$ となる Web ページのリンクはそのままに保ち、Web ページ群を構成していく。

3. 2 で構成した Web ページ群についてクラスタリングを行う。

4. 各クラスタ内で Web ページ群を組み合わせ、リンクの再構成を行う。

リンクの再構成の手順は、以下の通りである。

- ある Web ページ群 $S_{ci}$ を

$$S_{ci}(p_s^i, p_e^i)$$

と表す。ここで、$p_s^i$ は $S_{ci}$ の始点ページ、$p_e^i$ は $S_{ci}$ の終点ページを表す。ただし、Web ページ群が 1 ページのみのときは、

$$p_s^i = p_e^i$$

である。

- 各クラスタ内において、Web ページ群 $S_{ci}$ と $S_{cj}$ のページ群としての類似度 $sim(S_{ci}, S_{cj})$ を求め、一連の内容を伴った Web ページ系列 $d_{seq}$ を形成する。ここで、$sim(S_{ci}, S_{cj}) > v_{th}$ ならば、$sim(p_e^i, p_s^j), sim(p_e^j, p_s^i)$ を計算する。

$sim(p_e^i, p_s^j) > sim(p_e^j, p_s^i)$ のとき
$S_{ci} \rightarrow S_{cj}$ の順にリンクを張る。
$sim(p_e^i, p_s^j) < sim(p_e^j, p_s^i)$ のとき
$S_{cj} \rightarrow S_{ci}$ の順にリンクを張る。

- $sim(S_{ci}, S_{cj}) < v_{th}$ ならはじめに戻り、新たに別の Web ページ群を選択する。

- ユーザの入力した検索語 $q$ の重み $\boldsymbol{w}^q$ と Web ページ系列の重み $w^{d_{seq}}$ の類似度 $sim(w^q, w^{d_{seq}})$ を計算し、値の大きなものから検索結果とする。

## 4 まとめ

多数の検索エンジンの出力による負担を軽減するため、ユーザが所望のページにたどり着けるように、Web ページを巡回していくためのプランを生成するアルゴリズムを提案した。今後の課題として、Web ページ群構成のための閾値決定法と本提案方式の評価が挙げられる。

## 参考文献


[1] Lawrence Page, Sergey Brin, Rajeev Motwani, Terry Winograd, "The PageRank Citation Ranking: Bringing Order to the Web", online manuscript. http://www-db.stanford.edu/ backrub/pageranksub.ps

[2] http://www.google.com/

[3] http://www.yahoo.co.jp/

[4] http://www.goo.ne.jp/

[5] 山田誠二, 大澤幸生, "WWW での情報検索のためのナビゲーションプラニング", 人工知能学会誌, Vol.14, No.6, pp.1125-1133.

[6] "Authoritative sources in a hyperlinked environment", Proceeding of ACM-SIAM Symposium on Discrete Algorithms, 1998.

[7] Lieberman, H.: "Letizia: An Agent That Assists Web Browsing", In Proc. of IJCAI-95, pp.924-929.

[8] Joachims, T., et al. "WebWatcher: A Tour Guide for the World Wide Web booktitle", In Proc. of IJCAI-97, pp.770-775.