60-2

WebSkimming: WWW ページ群の動的要約による閲覧支援

角谷 和俊　正賀 信寛　上原 邦昭
神戸大学都市安全研究センター (工学部情報知能工学科)
E-Mail:{sumiya,shouga,uehara}@ai.cs.kobe-u.ac.jp

1 はじめに

一般に，Web ドキュメントは木構造あるいはグラフ構造をなすページ群から構成されている [1][2]. ユーザは，基本的に，一度に 1 ページのみ閲覧することが可能である．従って，複数のページがある場合には，Web ドキュメントをどのページから閲覧し，次にどのページに移るかはユーザが手動で決定していた．しかし，大量の Web ドキュメントを手動で閲覧することは非常に手間がかかる作業である．また，ユーザの要求として，Web ドキュメントの要約だけを見たいという場合があると考えられる．

本研究では，ある Web ドキュメントが与えられた場合に，そのドキュメントに含まれるページを，どの順番で，どのようにユーザに提示するかを自動的に生成する方式を提案する．すなわち，Web ドキュメントの要約 (skimming) を行う方法について検討する．また，要約した Web ページを直列化 (ストリーム化) し，ユーザに提示する方法について検討する．

2 Web ページ群の要約

本研究では，ある Web ドキュメントを特徴のある複数のページを含んで要約する方法について検討を行なう．ある Web ドキュメントに属する複数のページを閲覧する場合の例としては,「ある研究室のホームページ (Web ドキュメント) において,『趣味』についての記述があるページを閲覧したい」などの検索要求が考えられる．この場合，構成員のホームページでの中で，趣味に関するページが提示されるべきである．

例えば，図 1 の例では，網掛けされている Q, M, U の 3 つが特徴のページであるとすると少なくともこれら 3 つのページを含むページ群が，要約として抽出されるべきである．

2.1 共通親ページ

任意の複数ページから最短パスで辿ることのできるページを**共通親ページ**と呼ぶ．図 1 の例では，Q, M, U の共通親ページは E である．共通親ページを求める手順は以下の通りである：

(*Step* 1) 共通親ページが見つかっていない関連ページがあれば (*Step* 2) へ．なければ終了．

(*Step* 2) 各関連ページのリンク元ページを取り出す．

(*Step* 3) 得られたリンク元ページが自分が既にたどったことのあるページの場合はそのパスは削除する．得られたリンク元ページが自分以外が既にたどった，あるいは同時に訪れたページの場合はそのページが共通親ページとなる．(*Step* 1) へ．

2.2 パス生成

共通親ページと各関連ページとに関係するページを抽出するために，共通親ページから各関連ページへのパスを計算する．まず，ページ X からページ Y への**すべてのパス集合**を $X \overset{*}{\rightarrow} Y$ と表す．例えば，図 1 において，E から M へのパスの集合 $E \overset{*}{\rightarrow} M$ は $E \rightarrow H \rightarrow M$ と $E \rightarrow I \rightarrow M$ の 2 つのパスになる．

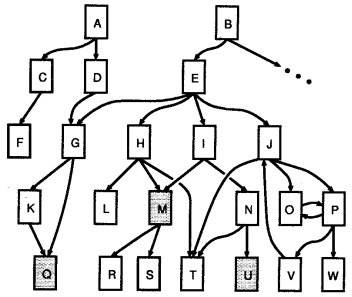

図 1: Web ドキュメントに含まれる関連ページの例

次に，求められたパス集合の中で，ページ X からページ Y へのパスとして最もふさわしいパス，すなわち最もコンテンツの内容を正確に表しているパスを 1 つ求める．このパスを**内容パス**と呼び，$X \rightsquigarrow Y$ と表す．

内容パスの生成手順は以下の通りである：(1) 終端ページにリンクしているページを取り出し，それぞれのページから出ているリンクの数の逆数を求める，(2) 求めた数値が一番大きいリンクを含むパスを選択する．これらを先頭ページまで繰り返す．図 1 の例では，$E \rightsquigarrow M$ は $E \rightarrow I \rightarrow M$ となる (図 2).

関連ページのすべてについて内容パスを求めると，あるドメインにおける関連ページを表す必要十分なパスの集合が求まる．これを**内容パス群**と呼ぶ．図 1 における関連ページ Q, M, U の内容パス群は，$E \rightarrow G \rightarrow Q, E \rightarrow I \rightarrow M, E \rightarrow I \rightarrow N \rightarrow U$ である．

[0] *WebSkimming*: Automatic Browsing for Web Pages based on Context Summarizing, Kazutoshi SUMIYA, Nobuhiro SHOUGA and Kuniaki UEHARA, Research Center for Urban Safety and Security, Kobe University

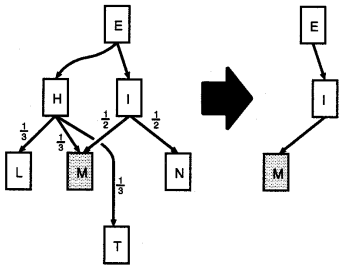

図 2: E から M への内容パス

2.3 Web ページのストリーム化

本節では，得られたパス群に含まれるページをストリーム (直列) 化する方法について述べる．ストリーム化の方針は以下の 2 点である:

1. ページ間の関連度を反映する
2. リンク構造を反映する

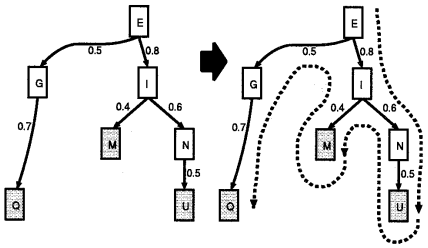

図 3: 類似度に基づくページ順序

ページ間の関連度として，$tf \cdot idf$ 法を用いて類似度を計算する．求められた類似度をパス群のパスに付与し，重み付けされた深さ優先探索によりページ順序を生成する．図 3 の左は，計算されたパスの類似度，右は生成されたページ順序である．この場合，$E \Rightarrow I \Rightarrow N \Rightarrow U \Rightarrow M \Rightarrow G \Rightarrow Q$ となる．

3 ページ・ストリームによる Web ドキュメントの呈示

本研究における呈示とは，ページ間の切り替え時にどのようにページを置き換えるかについてである．すなわち，PowePoint のスライドショーの切り替え効果[1] の様に，ページを切り替える場合に，そのページの Web ドキュメントにおける役割を反映した切り替え効果を自動的に選択する方法である．図 4 に切り替え効果の例としてワイプ (左側), ディゾルブ (右側) を示す．

ページには，そのページが含まれる Web ドキュメントにおける役割がある．例えば，たくさんのページへの栞の役目をしているページ，リンクアンカーの詳しい説明のページ，連続ページの中間ページなどである．これらの役割を，そのページの Web ドキュメントにおける特性によって計算する方式を提案する．ここでの特性とは，以下のファクターを指す．

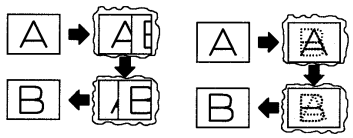

図 4: ページ切り替えの演出効果

- リンクされたページ数 (from-link)
- リンクするページの数 (to-link)
- ループリンクの有無 (loop-link)
- リンクされたページとリンクするページの類似度の差 (context-diff)

この 4 つのファクターを用いてページの役割を規定し，どのような切替え効果を選択するかを決定する．

4 おわりに

本論文では，ある Web ドキュメントが与えられた場合に，そのドキュメントに含まれるページを，どの順番で，どのようにユーザに呈示するかを自動的に生成する方式を提案した．また，呈示する際の演出効果についての考察を行なった．今後の課題として，実際の Web ドキュメントを用いた検証実験が挙げられる．

参考文献

[1] 永藤拓宏, 遠山元道. ページ群への分割を利用した www 検索エンジン. 電子情報通信学会データ工学ワークショップ (DEWS'98) 論文集, 1998.

[2] Yoshiaki Mizuuchi and Keishi Tajima. Finding context paths for web pages. In *Proc. of ACM Hypertext'99*, pp. 13–22, 1999.

[3] 清光英成, 田中克己. Web リンクの巡航に基づく動的なリンクの活性化とアクセス管理. アドバンスト・データベース・シンポジウム'99(ADBS'99), pp. 115–122. 情報処理学会, 1999.

[4] 品川徳秀, 北川博之. ユーザ視点に即した仮想 www ページの動的生成による閲覧支援. 情報処理学会研究会報告, 99-DBS-119, pp. 425–430.

[5] Site Cruise Theater.*http: //www.incx.nec.co.jp/sitecruise/*.

[6] 川崎成人, 水野浩三, 福岡秀幸. オンデマンド型 push 情報システムにおける番組提供とシナリオ記述. 情報処理学会第 55 回全国大会論文集 (4), pp. 349–350. 情報処理学会, 1997.

[7] 服部多栄子, 角谷和俊, 灘本明代, 草原真知子, 田中克己. 番組メタファーによる web ページの利用者適応型呈示方式. 情報処理学会研究会報告, 99-DBS-119, pp. 413–418.

[1] PowerPoint では，ワイプ，スライドイン，スライドアウト，ディゾルブ，フェードイン，およびフェードアウトなど約 40 種類の切替え効果が用意されている．