# 参照履歴を用いたWeb文書推薦方法の提案

5Z－5

織田 充　南 俊朗
（株）富士通研究所
{oda, minami}@flab.fujitsu.co.jp

## 1 はじめに

インターネットに代表される近年のネットワーク環境の広がりにつれて，ネットワークを通じた多種多様な情報サービスが提供され，利用者はそれらのサービスを容易に受けることができるようになった．特にWWW(World-Wide Web)を通じて利用できる情報量の増大には，目を見張るものがある．これらの大量情報の中から必要な情報を発見する困難さを軽減するために様々なサーチエンジンが提供されている．それらは，与えられたキーワードより，それを含むようなWeb文書の推薦を行う．しかし，それでもなお結果として膨大な数の文書が提示されたり，適切な文書が見つからないことも多く，有益な情報を持った文書を選別するための有効な推薦方式が求められている．

従来，推薦システムには，利用者が与えたキーワードなどの条件を基に，与えられた条件にあった内容を持った文書を推薦する内容による推薦方式や，利用者の好みなどの，いわゆるプロファイル情報の類似性に基づく推薦方式が多く用いられてきた．しかし，前者には，適切なキーワードや条件を与えることが困難であるという問題があり，一方後者には，多くのただのり(Free-Riding)利用者の存在等により質の高い適切なプロファイル情報が得るのが難しいという問題がある．

本稿では，このような状況を改善するために利用者の検索履歴を利用し，文書間の共起性を基にしたWeb文書の推薦方式を提案する．本方法は暗黙的方法で収集したプロファイル情報を利用するため，ただのりに関する問題を回避できる一方，利用者の実績に基づき候補の評価を行うため，必ずしも的確な条件を与えなくても，推薦された文書に対する選択を繰り返すことで，より適切な文書に到達できる可能性が高くなり，使えば使うほど，より有効な推薦が行われることが期待できる．

Using Referential History Data for Web-Documents Recommendation
Mitsuru Oda and Toshiro Minami
Netmedia Lab., Fujitsu Laboratories Ltd.
2-2-1 Momochihama, Sawara, Fukuoka 814-8588 Japan

## 2 文書の共起性

文書の共起性を説明する前に，まず，図1を用いて本方式の背景にある文書推薦機構を説明する．

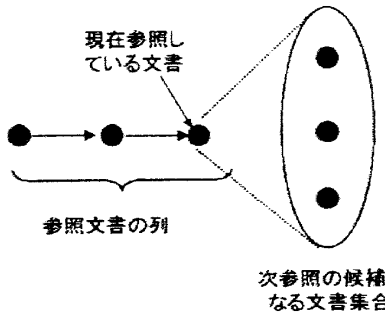



図 1: 文書推薦機構

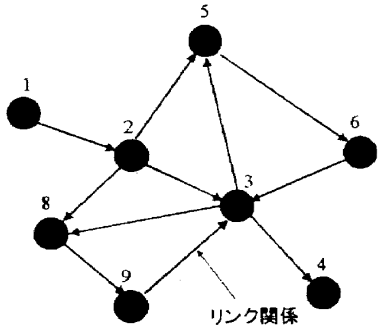
本図中央から左部分は，これまでの文書参照の経過を摸式的に表している．各黒丸は，参照された文書に対応しており，現在中央の文書が参照されている．図の右側の楕円内は，現在状況で選択可能な次の文書群を表している．このような状況において，過去どのWeb文書を経て現在の文書に至ったかの道筋は，何らかの意味で利用者の意図を反映しているものと考えられる．なぜならば，途中のそれぞれの文書において，その意図に従って最適と思われる文書へのリンクを選び現在に至っているからである．意図が大きく異なれば，異なったリンクを選ぶであろうし，他方，同じリンクを選んで現在の文書に至ったのであれば，元々の意図そのものが同じか，もしくは類似していたと考えられるからである．このような状況において，候補文書中のどれを次に参照すべき文書として選択するかが次の問題である．

本稿では“共起性”[2]という概念を利用し候補の推薦順位を定める方式を提案する．本概念は探索の向きを考慮する場合，しない場合等状況に応じた様々な定義が可能である．ここでは向きを考慮しない場合における，文書1と文書2の後に文書3が現れる共起率の求め方を示す．共起度を求めるためには，現在までに参照してきた文書集合$S$＝{文書1，文書2}と共起率を計算する対象となる文

書 $d$ = 文書3，そして，過去の成功検索履歴の多重集合 (multiset)$R$ が用いられる．まず，準備として履歴情報 $R$ に対して，現在の状況に対応した $S$ と交わりを持つ文書列のみを取り出した $R'$ を計算しておく．なお，$R'$ の要素である参照履歴列は集合として扱う．すなわち，$R' = [t|t$は$t' \in R$の集合化，$t \cap S \neq \emptyset]$ となる．この状態で，対象文書 $d$ の共起数を，$C = \sum_{t \in R' \text{ s.t. } d \in t} |t \cap S|$ と定める．共起度は，$S$ との交わりを持つ参照履歴列数に対する共起数の割合 $\frac{C}{|R'|}$ により定義する．

## 3 具体例

本節では，本稿の推薦方式を例を用いて具体的に説明する．文書および，それらの間のリンク及び過去の履歴情報は図2に示されるようになっているものとする．この状況構造の下，ある利用者が既に文書1，文書2，文書3の順にたどり，現在文書3を眺めているとする．図に示されるように，次にたどるべき文書の候補は，文書4，文書5，そして文書8である．




| 文書列の履歴情報 |
|---|
| 文書1，文書2，文書3，文書4 |
| 文書2，文書5，文書6，文書3 |
| 文書9，文書3，文書4 |
| 文書5，文書6 |

図 2: 文書間リンクと履歴情報の例

以下，文書は文書番号で表す．前節と同様に，これまでに参照した文書集合を現在の参照状況 $S$ とする．すなわち $S = \{1,2,3\}$ となる．過去の履歴情報 $R$ は個別の参照文書列の多重集合で表す：$R = [< 1,2,3,4 >, < 2,3,5,6 >, < 3,4,9 >, < 5,6 >]$.

次に，次参照のための候補となっている文書 $d \in \{4,5,8\}$ に対する共起度を計算する．$d = 4$ についての計算過程を示す．まず，$R$ の中から，現在状況 $S$ との交わりを持たない参照集合を削除する．その結果，$R' = [\{1,2,3,4\}, \{2,3,5,6\}, \{3,4,9\}]$ が得られる．次に，$d(=4)$ を含む各参照集合に対する，$S$ との共通要素の個数の総計を求める．本例の場合，総計は $(3+1=)4$ である．共起度は総計を $R'$ の位数（本例の場合は3）で割って求める．従って，参照状況 $S$ に対する文書4の共起度は $4/3 = 1.333...$ ということになる．他の $d$ に関しても同様に計算する．結果を，表1に示す．この結果より，本方式による文書の推薦は，文書4，文書5，文書8の順ということになる．

| 候補文書 | 評価値 |
|---|---|
| 文書4 | 1.3 |
| 文書5 | 0.7 |
| 文書8 | 0.0 |

表 1: 共起度の計算結果

## 4 まとめと今後の課題

本稿では，文書間の共起性を利用したWeb文書推薦方式を提案した．本方式の実効性は，我々が目的の文書を探索する過程において，本当に高い共起性があるのか，また，あるとすれば，どのような計算式でそれを的確に把握できるかにかかっている．今後，実際のWeb文書に対する実験のためのプロトタイプシステムを実装し，共起性定義の実効性を確認する．また，Web文書に限らず，エージェント社会 [1] 等の様々な場面において逐次的な探索は行われており，本稿方式の効果の高い適用領域の見究めも今後の重要課題である．

## 参考文献

[1] 南俊朗，有馬淳，織田充，大谷武． エージェントと仮想社会， 人工現実感に関する基礎的研究（九州地区）シンポジウム，重点領域研究「人工現実感」総括班， pp.59-62, 1997.

[2] 南 俊朗，織田 充． 関連度を用いたWeb文書のナビゲーション． マルチメディア通信と分散処理研究会，情報処理学会, 2月 1998.