システムソフトウェアと
オペレーティング・システム 77－2
マルチメディア通信と分散処理 87－2
（1998. 2. 26）

関連度を用いたWeb文書のナビゲーション

南 俊朗　　織田 充
（株）富士通研究所 ネットメディア研究センター
〒814-8588 福岡市早良区百道浜2-2-1 富士通九州R & Dセンター
{minami, oda}@flab.fujitsu.co.jp

サーチエンジンの利用により，World-Wide Web上に提供される膨大な量の文書情報から，利用者にとって有益な文書を絞り込み利用することができる．しかし，それでもなお数多くの候補が選ばれ，その中から本当に有効な文書を捜し当てる困難がある．本稿では，この検索作業を軽減させる一方式として，それまでに参照された文書との関連性の高い文書を推薦することで，利用者の求めている文書へとナビゲートする方式を提案する．また，学習教材を例として順序性のある状況でのナビゲーションに関する考察と，関連度の計算方法を示す．更に，エージェント社会における分散型ナビゲーションへの適用について述べる．

Navigation through Web Documents using Co-Relations of Documents

Toshiro Minami　　Mitsuru Oda
Netmedia Lab., Fujitsu Laboratories Ltd.
2-2-1 Momochihama, Sawara, Fukuoka 814-8588 Japan

It is a difficult task for users to find appropriate web documents among a lot of documents retrieved by the search engines, where searches are performed with the keywords given by the users. In order to relieve this situation, this paper proposes a new method of navigating the users by recommending documents based on the co-relation of web documents. It also proposes a co-relation function by taking learning materials as an example and considers the situation where the naviagation ordering is important. Further, it shows that this method can be applied also to the distributed naviation in the agent society.

1 はじめに

近年，インターネットに代表されるネットワーク環境の整備が進み，Web文書に見られるように，これまではそれぞれのサイトに独立して存在していたデータベースが，ネットワークを通じて相互に結合されるようになり，この状況変化を受けて，多くの利用者に多種多様な情報サービスがネットワーク上で提供されるようになった．このような環境における情報検索を手助けするためのサービスとして，Yahoo[6]を始め様々なサーチエンジンがインターネット上に提供され，利用者の与えたキーワードより，候補となるWeb文書を探し出し提示してくれる．ところが，元となるネットワーク上の情報量が膨大であるため，サーチエンジンから返される文書群もまた膨大になる場合が多い．このため利用者にとって，有益な情報を含む文書を選別することが重要な問題となっている．

サーチエンジンにおいても，単に全文検索による提示だけではなく，情報の有益さに関する評価(Rating)を行い，その結果に基づいた提示を行っている．しかし，多くの場合その評価の基準は，キーとなる用語が含まれる頻度等の表面的な情報によっており，利用者の意図との関連が必ずしも強いとはいえない．その結果，選別された結果に不要な情報が多く含まれることも多く，より効果のある選別法が待たれている．

従来より，推薦システムでは，サーチエンジンに見られるように利用者が与えたキーワードなどの条件を基に文書を推薦する内容による推薦方式，及び，利用者の好みなどのプロファイル情報の類似性に基づく推薦方式が代表的である[4]．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る．また，後者にも，適切なプロファイル情報が得られるならば，それに対応した質の推薦が得られるものと期待できるが，現実には，いわゆる，ただのり(Free-Riding)の利用者も多く，十分な質を確保できるだけの情報を集めるのが困難である．

本稿では，このような状況を改善する方法として，利用者の検索履歴を利用し，文書間の関連性を基にしたWeb文書の推薦方式を提案する．本方式は，暗黙的なプロファイル情報の収集を行うため，ただのり問題を回避できる一方，利用者の実績に基づいた候補の評価を行う方式であるため，かなら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まり，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．

以下，本稿は次のように構成される．第2節では，目的とする文書／情報の探索スタイルについて分類を行い，次の第3節では特に推薦によってWeb文書を探索するナビゲーションモデルに関する考察を進める．第4節では，順序を考慮した探索過程での文書間の関連度の定義を与える．ここで述べた探索方式は，Webブラウザによるナビゲーションのみならず，エージェント社会における分散型情報探索にも有効であるものと考えられる．第5節では，そのような状況への適用法に関する説明を行う．最後に第6節において，全体の総括を行い，今後の課題を述べる．

2 Web文書探索機構

本節では，まず，準備として，検索と探索の性格の違いを論じ，本稿で用いられる探索という概念を明確にする．その後，いくつかの検索スタイルを挙げ，それらの特徴に関する考察を行い，推薦機構の特徴を論じる．

2.1 検索と探索

我々の興味は，我々の抱えている問題を解決するために役立つ情報を含んでいるWeb文書を効率良く発見し，必要な情報を入手する仕組みにある．本稿では，情報の入手方法を，入手する手続きの性格の違いによって検索と探索に区別して考える．

検索は，基本的に情報の所在ははっきり分かっている場合に用いる．例えば，データベースを検索する場合，そのデータベースにどのようなデータが蓄積されているか予め分かっている．利用者は，自分の欲しい情報を何らかの方法で記述し，この記述に基づいて検索機構がデータベースから要求された情報を取り出す．このような性格の情報入手を検索と呼ぶ．

一方，探索の場合，情報の所在が必ずしも予め分かるとは限らない，そもそも，そのような情報が存在するかどうかも事前にはっきりしない．更に，探索者自身が自分の求めている情報を明確に把握していないことさえありうる．このような状況での情報発見を探索と呼ぶ．

2.2 探索スタイル

本節では，3種類の探索スタイルを取り上げ，それらの間の比較を通して，後で特に取り上げられる推薦スタイルでの探索を特徴づける．

絞り込み型探索 これは，サーチエンジンを用いた探索において良く用いられている方式である．利用者は，まず自分の求めている文書を表現すると思われるキーワードを与える．ここに，キーワードという用語は，キーとなる単一の語のみでなく，もっと広い意味で用いる．例えば，実際は複数の語で与えられる場合も含めてキーワードと呼ぶ．また，同じ複数の語に対しても，それら全てを含む文書であるとか，いずれかを含む文書であるとか解釈も様々でありうる．以下，それらを一括してキーワードと呼ぶことにする．

システムはキーワードに関連すると思われる文書群を利用者に提示する．通常，最初に与えられたキーワードにより得られる文書は大量であり，また利用者の意図とは異なる文書も含まれる．キーワードを追加することで目的の文書を絞り，選択される文書数を少なくしていく．

本方式は比較的単純であり分かりやすいものであるが，効果的な絞り込みを行うための適切なキーワードを思いつくのが難しいという問題点がある．また，キーワードが適切でないために，絞り込まれたリストから，重要な文書が抜け落ちてしまうことも問題である．

嗜好の類似による探索　本方式は，Social Filtering もしくは Collaborative Filtering と呼ばれる方式であり，嗜好の類似した別の利用者の好みや評価の高いものを利用者に推薦するというものである．例えば，Ringo[5] に見られるように音楽などの感性に関するものへの適用や，いわゆるネットサーフィンのように明確な目的を持たず，興味を引くいろいろな文書を見てあるくという目的には一定の効果が期待できるものの，本稿で想定しているような，現在利用者が探している文書を探索するための仕組みとしては十分ではない．

推薦による探索　本方式は，文書から文書へと渡り歩く中で，多少なりとも目的の文書に接近し，ゆくゆくは目的の文書に到達しようという探索のスタイルである．システムは，利用者の渡り歩いた履歴を参考に，目的となる文書への方向性を推測し，推薦文書群として利用者に提示する．利用者は，提示された文書リストの中から，自分が探索している文書へ到達する可能性が高いと思われる文書を選択する．このような過程の繰り返しの中から，利用者の目的が明確化され，最終的には目的の文書へ到達することを目指す．通常，Web 文書の場合，リンク情報により，次に辿るべき文書群が示されている．

本稿では，システム自体が利用者の目的を推測し，その推測に基づき推薦度の高い文書で，元の文書のリンクに現れないものについて，リンクを追加して利用者に提示することをも含めた想定を行う．このように追加することで，利用者を目的の文書へ的確にナビゲートすることが期待できる．

本方式は，探索する対象に対するある程度のイメージはあるものの明確に規定するのが困難である場合に特に適している．

3　推薦によるナビゲーション

本節では，推薦による探索モデルに基づくナビゲーション過程のモデル化を行い，それに基づき関連性を利用した推薦を行うための要件を分析する．

3.1　推薦によるナビゲーション過程のモデル

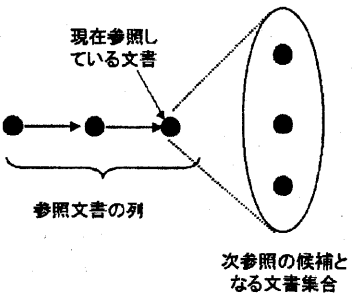

図 1: 文書推薦機構

図 1に，我々が想定している Web 文書の推薦機構のモデルを示す．本図中央から左部分は，これまでにどのような順で文書が参照されたかの経過を摸式的に表している．各丸印は，参照された文書を示しており，現在図中央の文書が参照されている状況が表されている．図の右側部分の楕円内は，現在状況で選択候補となる文書群を表している．Web 文書の場合，通常次に参照される候補文書は現在参照している文書からリンクを張られた文書であるが，本稿では，直接リンクを張られていない文書をもシステムが推薦する場合をも想定している事に注意して欲しい．

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である参照文書列は，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．

このような理解を基に，候補文書集合に属す各文書の推薦優先度をどのように定めるかが重要な課題であり，以下優先度を定めるための要件について議論する．

3.2 現在状況の定義

図1によれば，現在状況の表現として，これまでに参照されてきた文書列（以下Sと表す．）を採用するのが妥当である．しかし，Sに現れる文書が全て次に参照される文書の決定にとって重要な訳ではない．

例えば，国語，数学，社会，理科の4科目の平均を求める問題を取り上げよう．平均値を求めるためには，それぞれの教科の点数が必要である．そのためには，これら4教科全ての点数が必要である．これまでに，国語，数学，社会の点数が分かったとする，これまでの探索過程は，一般には，何らかの別情報が含まれているであろう．例えば，$D1$ → 数学 → $D2$ → $D3$ → 国語 → 社会 → $D4$ といった具合である．この状況の場合，

$$S =< D1, \text{数学}, D2, D3, \text{国語}, \text{社会}, D4 >$$

となる．この際，関連性を定義する上で重要なのは，すでに数学，国語，社会の点数が分かっているということであり，その他の情報を含んだ文書1，2，3，4に関しては重要ではない．従って，本例の状況に見られるように状況Sは，それ自体が全体として重要な訳ではなく，それらの部分情報が重要であるということになる．

もちろん，S全体として意味のある状況も存在するものとは，考えられるが，利用者がナビゲーション情報を求めている状況を考えると，一般に，状況Sが全体として次候補を選定するのに重要である場合よりも，そこに含まれるある部分が決定に重要な役割を演ずる場合が圧倒的に多いと思われる．

以上の考察により，以下の関連性の定義においては，状況S全体ではなく，その一部分との関連性を用いることにする．

3.3 参照順序の考慮

上記のモデルでも見たように，文書の推薦に基づく参照経過は一般に参照列としての構造を持つ．上記の平均値計算の例の場合は，結果としてどのような文書が参照されたかが重要である．一方，計算機による教育 (CAL/CAI) システムにおける学習過程を表すコースウェアの場合，学習する順序関係は重要である．この例の場合，各学習内容は，文書に対応し，ある学習の次に何の学習を選ぶかがリンクに対応する．このようなコースウェアの場合，結果としてはある学習内容の集合を学ぶことが目標となるが，個々の内容選択の際は，それまでに学んだ内容に依存して次に学ぶべき内容が大きく異なる．

例えば，小学校の算数教材の学習の場合，足し算を学んだ後に引き算を学んだり，かけ算を学んだりできる．先に，かけ算を学び，その後，足し算を学ぶということは決して起こらない．一方，かけ算と引き算のどちらを先に学ぶかに関しては，選択の余地がある．引き算を知らなくてもかけ算を学ぶことは可能であるし，逆にかけ算を知らなくても引き算を学ぶことができる．従って，Sの中に足し算が含まれているならば，引き算やかけ算は推薦することができる．

割り算を学ぶためには，かけ算と引き算を学んでいる必要がある．従って，状況Sにかけ算と引き算が含まれているならば，次の学習候補として割算を推薦することができる．ここで重要なのは，Sの中で，かけ算が引き算よりも先に現れたのか，それとも後に現れたのかは問題ではないことである．結果として両者を学習してあれば，割り算へ進むことができる．

この例に見られるように，次にどのような対象を推薦するかに関しては強い順序が必要とされる場合においても，一旦参照されてしまった対象に関しては，結果としてそれらが参照されてしまっているということだけが重要であり，既参照の状況Sの中での順序は重要ではない．従って，次節において，関連度を定義する際，参照列情報列Sの集合としての属性に依存した定義を与えることが妥当である．

4 関連度の定義

前節で述べたような状況において，次参照の候補文書集合に対する推薦の優先度を定めるた

めに文書間の関連性を利用する．関連度とは，現在の状況，すなわち，これまでに参照されて来た文書群に対する次の参照候補となる文書の関連性の程度のこととする．

このような考えの元であり，第1近似となる概念は確率論における条件付確率もしくは事後確率である．これまでの参照文書履歴を S とするとき，対象文書 t に対する条件付確率 $P(t/S)$ を関連性の程度である関連度の定義とすることも可能である．しかし，有効性の高い関連度の定義を行うためには，前節で述べた，文書推薦のモデルに対するもっと精密な定義が求められる．

まず，準備として，以下の説明のために必要な数学概念の定義を行う．

列は "⟨" と "⟩" で括って表す．例えば，文書列 $a \to b \to c$ は，$\langle a, b, c \rangle$ と表される．空列は $\langle\rangle$ と表される．列 S_1 と列 S_2 の合成は，$S_1 S_2$ と並置する．すなわち，$\langle s_1, s_2, \ldots, s_n \rangle \langle s_{n+1}, s_{n+2}, \ldots, s_m \rangle = \langle s_1, s_2, \ldots, s_n, s_{n+1}, s_{n+2}, \ldots, s_m \rangle$ となる．列 S' が要素 t を含む時，S' の，要素 t による最右切断の左断片 S'/t を次の命題の T_1 として定める．

命題 S' を列とする．$t \in S'$ のとき，次を満たす，列 T_1, T_2 が常に唯一存在する．

$$T_1 \langle t \rangle T_2 = S', \text{ かつ } t \notin T_2$$

（証明略）

多重集合は "[" と "]" とで括って表す．また，集合，多重集合，そして列の要素数を $|\cdot|$ で表す．その他，多重集合や列に対して集合演算を施す場合，それぞれを一旦集合に変換し，その後，集合演算を行うものとする．

以下，参照文書列を S と表し，次に参照される候補文書の集合を D と表す．また個々の候補文書を t で示す．過去の参照履歴の記録は，参照文書列の多重集合 R で表される．

以上の準備の下，順序を考慮した場合の関連度を定める．簡単のために，$S = \langle$文書1$\rangle$ の後に $t =$ 文書2 が現れる場合の図2を参照しつつ説明する．本図は，文書の探索履歴の内，文書1を参照した場合を選び出した状況を摸式的に表している．

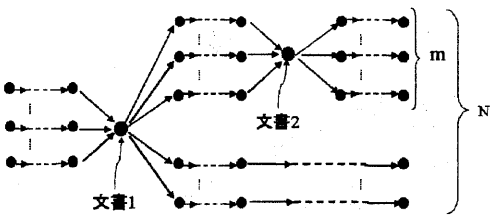

図 2: 関連性の基本原理

まず，履歴データ R の中から現在状況 S に関連する参照列 R' を取り出す．すなわち，$R' = [S' \in R \mid S \cap S' \neq \emptyset]$ と定める．図2に示されるように，$N = |R'|$ とする．

関連度を次の式で定義する．

$$\frac{1}{N} \sum_{S' \in R' \text{ s.t. } t \in S'} |S \cap (S'/t)|$$

なお，m を，R' の中で，文書1の参照後に文書2が参照された場合の数とすると，図2のように S が1個の要素のみを含む場合，$\sum_{S' \in R' \text{ s.t. } t \in S'} |S \cap (S'/t)| = m$ となるため，関連度は m/N であり，文書1の参照の後に文書2が参照される条件付き確率と一致する．

5 分散エージェント推薦方式への適用

本節では，図1で示された文書推薦機構の応用として，元のモデルにおける文書を，ネットワーク上に分散されたエージェントと読み代えることで，本モデルをネットワーク上に分散配置されたエージェントのつくる社会的組織 [1] による協調的な情報資源の検索サービス提供モデルへ適用する方法を紹介する．

図3は，ネットワーク上に分散配置されたエージェント群が協調して利用者の要求する情報資源を探索する様子を表している．図中の丸印はエージェントに，また，矢印は，リンク関係に対応している．

情報資源探索は次のように行われる．まず，左上にいる情報利用者が，自分の欲しい情報を身近なエージェントに伝える．エージェントは，利用者の要求に合う情報を知っているならば，それを利用者に返す．一般には，利用者からの要求は，1つの情報を発見するだけで良い場合や，探せる限り多くの文書を発見する場合など，様々である．前者の1つの文書発見の場合は，

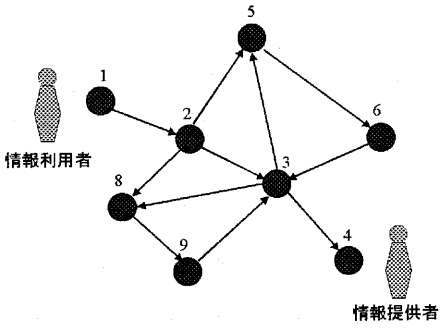

図 3: エージェント社会による分散型文書推薦

エージェントが文書を見つけた時点で利用者の依頼は解決されるが，後者の場合は，エージェントは，その近隣のエージェントへも利用者の要求を伝達する必要がある．

利用者の要求は，エージェントのネットワークを通じて伝えられ，最終的にその要求に合った情報の提供者もしくは，その提供者が，その情報を持っているという情報を持つエージェントに到達する．ここで，提供者に関する情報は，予め広告として周辺のエージェントに伝えられたあったものとする．利用者からの要求にマッチした提供者情報に出会ったことで，目的の文書や情報に関する情報は，同じくエージェント間の情報伝達路を利用して要求元の情報利用者に伝えられる．

前節までに説明した機構をこのようなエージェントモデルに当てはめる．各エージェントは，自分の関わった探索の結果を探索履歴情報として蓄える．エージェントに伝えられる要求には，それがどのような経路を通って伝えられたかの情報を加えられる．要求を受けとったエージェントは，その要求のデータに記入された経路情報を前節までの参照文書列に相当するものと解釈する．また，エージェントは，自分が隣接エージェントとして把握しているエージェント群の中から，状況を表す経路情報と，自分が蓄積している過去の履歴情報を用いて，どの隣接エージェントに要求を転送すると解決できる可能性が高いかを判断する．この判断の際に，前節で定義された関連度を用いた推薦法を適用することで，情報を伝達するのに有効性の高いエージェントを選択することができる．

6 まとめと今後の課題

本稿では，既に参照された文書群との関連度を利用した Web 文書推薦方式に関し，ナビゲーションモデルの説明を行い，またコースウェアを例に，順序性のある状況における関連度の定義例を紹介した．

本方式の実効性は，我々が対象とする問題領域で目的の文書を探索する過程において，文書間の関連性が本当に高いのか，あるとすれば，その計算式の適切さにかかっている．従って，プロトタイプシステムを実装し Web 文書に対する実験を行い，関連度利用の有効性を実際に確認することがもっとも重要な今後の課題である．また，どのような実例において関連性の高い探索が行われているかの，適用領域の見究めも重要であり，今後検討を行う．特に，第 5 で述べたようにエージェント社会 [1] における問題解決機構，例えば，口コミエージェント方式 [2, 3] に関連度による推薦機構を適用することで，より効果の高いナビゲーションの実現が期待できる．

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] 大谷 武，南 俊朗：口コミによる情報資源探索，第 6 回マルチ・エージェントと協調計算ワークショップ (MACC), 日本ソフトウェア科学会，12 月 1997.（掲載予定）

[3] 大谷 武，南 俊朗：エージェント社会における資源管理，電子情報通信学会「人工知能と知識処理」，情報処理学会「知能と複雑系」合同研究会報告，1 月 1998.

[4] P. Resnick and H. R. Varian (Guest Eds.): Recommender Systems, CACM, Vol. 40, No. 3, pp.56-89, March 1997.

[5] U. Shardanand: Social Information Filtering for Music Recommendation, TR-94-04, Learning and Common Sense Group, MIT Media Lab., 1994.

[6] Yahoo: `http://www.yahoo.co.jp/`, `http://www.yahoo.com/`.