トピカモデルにおけるユーザアクセス経路を利用した関連文書の検索

浅妻理 田中譲 藤間淳
北海道大学大学院情報科学研究科
コンピュータサイエンス専攻知識メディア研究室

1 はじめに

膨大な量のウェブ文書の中からユーザが必要とする情報を抽出するために、様々な検索モデルがある．しかしながら，これら既存の検索技術はコンテンツが類似したものを抽出するのには適しているが，ユーザのアクセスコンテキストが類似したものを抜き出すのには適していない．

ウェブ上でのユーザの操作は固定的になりがちであり，そのため自ら興味のある内容の文書を見逃しているかもしれない．例えば芸能人とスポーツ選手の結婚情報を取得する場合，スポーツ選手に興味のある人はポータルサイトのニュース⇒スポーツのカテゴリから情報を取得するであろうし，芸能人に興味のある人はニュース⇒芸能のカテゴリから情報を検索するであろう．今までその芸能人に興味が無かった人が，その結婚を通して興味を持ったとき，その芸能人について後者の検索経路を使うことにより，詳しい情報を取得できる．また辿るリンクは同じでも，利用しているポータルサイトが違えば異なる情報を獲得できる．

そこで本論文では，類似したアクセス経路を通ってたどり着けるウェブ文書を，類似したコンテキストを持つ関連文書と定義する．ユーザが任意のウェブ文書内に他のウェブ文書を配置し，その文書内で組織化できるトピカモデルという情報管理アクセスモデルを基盤として用いて，そのモデル内でのユーザのアクセス経路情報を任意のウェブ文書に付加する．その情報を用いて関連文書の検索を実現可能とするための枠組みを提案する．

2 データ管理アクセスモデルの概要

2.1 トピカモデル

トピカモデル [1] とはオブジェクトの量と多様性が増すなかで，それらの管理手段として北海道大学の田中が提唱したモデルである．現実世界におけるカタログなどの消費者に対する商品アクセス支援を参考にし，文書にオブジェクトを格納する場所 (トポス) を設けることによって，複数のオブジェクトを関連付けて管理することを可能にしている．トポスが設置された文書をトピカ文書といい，そのなかに格納されたオブジェクトはトピカテーブルと呼ばれる関係表に関連付けられた状態で配置される．

ユーザはトピカ文書を介して，トポスにアクセスすることができ，トポスに制約を与えること（Fix）によって，アクセス可能なオブジェクトを絞り込むことも可能である（図 1）．

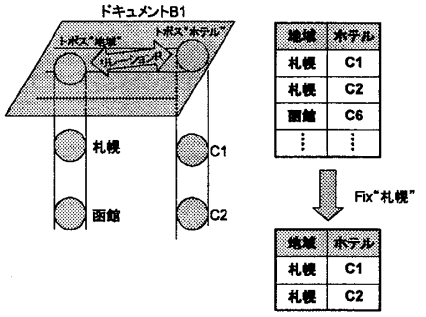

図 1: トピカ文書へのオブジェクトの格納

2.2 トピカ文書の作成

既存のウェブ文書にトポスを設置しトピカ文書として扱う．トポスの定義情報は，アノテーションサーバに保存される．また他のウェブ文書をトポスに格納していくことにより属性名をトポス名，要素をウェブ文書とする関係表が構成される．それにより，一つのウェブ文書上でトポスに格納された文書同士が相互に関連付けられた状態となる．各々のトポスには，その名前や場所にふさわしい文書やテキストが格納されているとする．例えば図 2 において，“観光”トポスに格納されている文書 B1～B3 は観光に関する記事であり，“地域”トポスに格納されている文書は地域名を表すテキストである．

2.3 トピカ文書へのアクセス

図 3 でトピカモデルにおける情報取得の例を示す．ユーザはまず A1 の“観光”トポスに格納されているウェブ文書 B1，B2 のなから B1 を選ぶ，B1 のウェ

Retrieval of relational documents based on users'access paths in Topica model.
Osamu Asazuma , Yuzuru Tanaka , Jun Fujima
Meme Media Laboratory,Hokkaido University
N13W8,kita-ku,Sapporo,060 8628,Japan

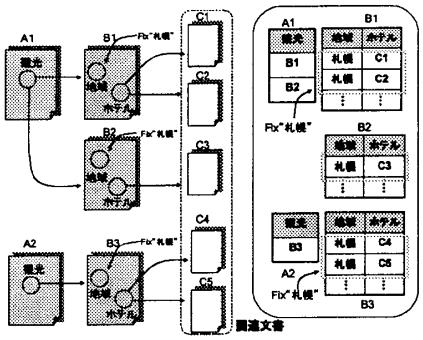

図 2: トピカモデルにおける情報取得

ブ文書を開いたときに同時にアノテーションサーバから B1 についてのトポスの情報やトピカテーブルの情報がダウンロードされて表示される．その後 B1 において“地域”トポスに格納されている文字列「札幌」を使い Fix を行う．トピカテーブルにおいて関連のない文書が除かれて“ホテル”トポスの中には C1，C2 が選択可能な要素として残る．

3 経路情報に基づいた検索

3.1 経路情報

一つの文書にアクセスするたびユーザの経路情報は生成される．この経路情報は，アノテーションサーバに保存されている経路と異なれば保存され，一緒であれば保存されない．これによって一つの文書は複数の経路情報を持つことになる．ユーザはある興味ある文書にたどり着いたとき，その文書に付与されている経路情報を見ることができ自分が辿ってきたのと異なる経路情報に基づく関連文書も検索することができる．

トピカ文書上で行える操作は，(1) どのオブジェクトで Fix したか (2) どのトポスのオブジェクトを選択したかの 2 つに限定される．この操作履歴の組み合わせを，ユーザのアクセス経路情報としてトピカ文書に付加する．図 2 において A1 から C1 まで到達した経路情報は図 3 のように表される．

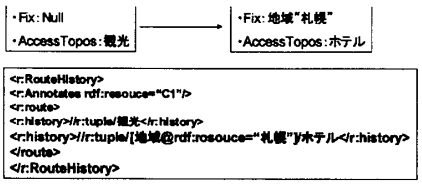

図 3: アクセス経路情報の RDF 表現

3.2 関連文書

C1〜C5 は全て同じ経路情報を保持している．その内容は札幌についての観光ホテルの内容などを表すウェブ文書であると考えられる．また最初のアクセストポスが観光ではなくビジネスである場合，札幌に関するホテルビジネスの内容のウェブ文書にアクセスすることになると考えられる．このように同じ経路を辿ってたどり着ける文書は似た，もしくは関連した内容を表している．よって，本論文では同じ経路情報を持つ文書は似たような内容を表す関連文書と定義している．

3.3 関連文書取得の流れ

文書の内容などからキーワード検索が難しい場合，ユーザはウェブ上でトピカモデルを構成することを可能にしたトピカブラウザを利用することができる．トピカブラウザはウェブサーバからウェブ文書を取得すると同時にアノテーションサーバからそのウェブ文書におけるトポスの定義，トピカテーブル，アクセス経路情報を取得する．その後いくつかの文書のトポスを辿り文書を表示させると文書についての経路情報がそのつど生成され，アノテーションサーバに保存される．最終的に獲得した文書の経路情報をアノテーションサーバにクエリとして送ると，関連文書として同じ経路情報を持つ文書が表示される．こうしてユーザはキーワード検索を使うこと無く関連文書を取得することができる（図 4）．

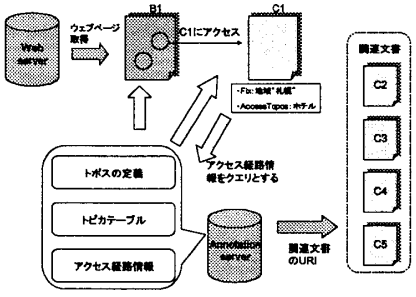

図 4: 関連文書取得の全体図

4 おわりに

本論文では，トピカモデルという情報管理のモデルを用いてウェブ上の文書を組織化し，その中から関連文書を検索する手法を示した．従来の方法とは違いウェブ文書の内容からの文書検索ではなくアクセス経路に基づく検索であるのでユーザが検索すべき内容を語句で表現できない場合でも任意のウェブ文書について関連する情報を獲得することができるようになる．

参考文献

[1] Yuzuru Tanaka. Meme Media and Meme Market Architectures: Knowledge Media for Editing, Distributing, and Managing Intellectual Resources. July 2003, Wiley-IEEE Press.