文書の話題構造を利用した意味的適合フィードバック機構

6W−6

平田 陽一† 田中 克己‡
† 神戸大学大学院自然科学研究科 ‡ 京都大学大学院情報学研究科

1 はじめに

本研究では，Web文書の話題構造を自動抽出し，その話題構造を比較することによりWeb文書間の意味的な関係を発見する方式と，そのような意味的関係を持つ文書を呈示してくれる意味的適合フィードバック機構を提案する．話題構造は主題的・内容的な単語群の二対グラフで表される．従来の適合フィードバックはWeb文書に対する評価を「良い」「悪い」だけで行うが，本研究では「今見ているページはいいのだが，もっとこの部分について詳細なページが欲しい」「もっと簡潔なページが欲しい」といったような意味的な評価を許容することから意味的適合フィードバックと呼ぶ．

本手法の利点は3つある．ユーザに検索結果内の主題・内容キーワードを呈示することによって検索結果全体をユーザに見せることができること，またキーワードや話題グラフを見せることによってユーザの意味的適合フィードバックにおける質問生成を助けること，またユーザの応答によって動的にWeb文書やシステムが呈示するキーワード群が変化することである．

本稿では，まず話題構造の抽出方法とそれによる意味的関係の発見方式を述べる．次に話題構造を利用した意味的適合フィードバック機構を提案する．

2 話題構造抽出と意味的関係発見

情報検索システムの問題点は，検索精度を高めるために大多数のユーザが必要とすると思われる情報を多く含んでいるページを適合とみなし，そのためにユーザの個人的な要求を満足させるページを発見するのが難しいことである．このような問題を解決するためには結果に対するフィードバックをユーザ自身が行う方法が有効であるが，従来の適合フィードバックにおいて導き出されるのは類似・非類似という2つの関係を持つページのみであり，求めるページがサンプルに対してどのように異なるのかといった，より複雑な関係を持つページを導き出すことはできず，またそのようなフィードバックを可能にするための複雑な関係をページ間に定義することのできるシステムも存在しなかった．そこで我々はさらに複雑なフィードバックを可能とするためにこの情報間関係の定義に関して次の方法を提案した[1][2]．Web文書を複数の話題で構成された集合体であると考え，文書が幾つの話題で構成されているのか，また各話題はどのように構成されているのかといった情報を単語の出現密度分布[3]と共出現傾向を利用した二対グラフ(話題グラフ)を用いて表現し，その話題グラフを比較・検討することで文書間の複雑な関係を発見していく．

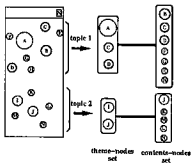

図1: Topic graph

2.1 話題グラフ

1つの話題は複数のキーワードから構成されている．キーワードには話題の主な特徴を表す主題的なキーワードと主題のキーワードを解説するための内容的なキーワードの2種類がある．話題間の関係を定義するにはこの2種類のキーワードをそれぞれ比較する必要がある．

2.2 主題キーワード

主題キーワードの抽出には，単語の出現密度を利用する．単語の出現密度とは，文書中での任意の位置を中心とした，ある範囲内の単語の頻度と位置的な情報を元にして算出する数値である．さて，ある話題をあらわす主題キーワードとは各話題内において広範囲に頻繁に用いられていると考えられる．そこである範囲内のすべてにおいて出現密度がある一定の閾値を超えている単語を主題キーワードとする．

2.3 内容キーワード

主題キーワードに深い関連がある単語はその主題キーワードが表す話題の要素であると考え，内容キーワードとする．ここでは次の2種類の関連を調べている．

- 位置依存関連

  単語の文書中における記述位置に基づいた関連である．ある2単語が文書中で近くに記述されている場合その2単語には強い関連があると考えられる．これは出現密度を使って求める．

- 非位置依存関連

Semantic relevance feedback uses topic structures of web documents
†Yoichi Hirata, Graduate School of Science and Technology, Kobe University. ‡Katsumi Tanaka, School of Informatics, Kyoto University.

単語の文書中での記述位置に関係ない関連である．文書中では近くに記述されていないものでも密接な関係を持つ傾向のある単語の組が当てはまる．これは共出現傾向を調べて求める．

2.4 意味的関係の発見

話題グラフの比較によってWeb文書間の意味的関係を発見する．例えば，図2の2つのグラフは「詳細・簡潔」な関係にある．

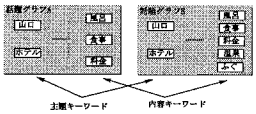

図 2: Semantic relationships

3 意味的適合フィードバック機構

上に述べた話題グラフを利用した意味的適合フィードバック機構について述べる．

まずシステムの流れを述べる．最初にユーザに主題・内容キーワードを入力してもらう．次にサーバはこれを検索キーワードとして既存検索エンジンで検索し，検索結果リストを獲得する．そして獲得した検索結果の話題グラフを作成し，ユーザの入力した主題・内容キーワードを含むWebページのみを検索結果としてユーザに返す．(図3上) フィードバックの際には，ユーザの要求をサーバが読み込み，検索結果から適当な話題グラフを持つWeb文書を選び出す．ここでのフィードバックによって検索結果の再評価，再ランキングが行われている．(図3下)

次にシステムの表示画面について述べる．(図4参照) 左上は検索結果のWebページである．その右が検索結果全体の主題・内容キーワードの表示欄，Webページの下のグラフは今表示されているWebページの話題グラフである．話題グラフ表示欄の右は「フィードバック」ボタンである．「詳細」「簡潔」「別話題」といったものがある．

例えば，最初の検索キーワードを「ホテル・山口」とすると，図4のような画面が現れる．そして「詳細」ボタンを押すと，図2の右の話題グラフを持つWebページが，また「別話題」ボタンを押すと主題・内容キーワードがそれぞれ異なる話題グラフ，例えば主題キーワードが「きらら博・山口」で内容キーワードが「場所・期間・アクセス」のような話題グラフを持つWebページが新たに表示される．

この機構には次のような特徴がある．それは検索結果の全体像が把握できること，ユーザの質問生成の補助機能があること，ユーザの応答によってWebページや主題・内容キーワードや話題グラフといったものが動的に変化する視覚的娯楽要素があること，そして従来の適合フィードバックではできなかった意味的な要求を満たすことのできる意味的適合フィードバックが可能になっていることである．

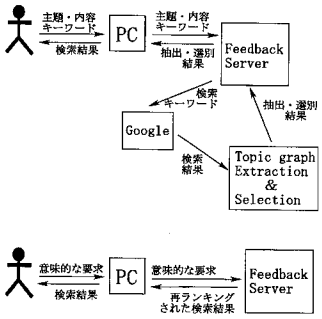

図 3: System

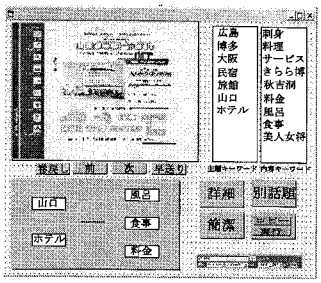

図 4: Semantic relevance feedback

4 おわりに

本稿では出現密度分布と共出現傾向からWeb文書の話題構造を抽出し，意味的な関係を含ませた質問を返すことのできる意味的適合フィードバック機構を提案した．今後の課題はシステムの実装，ユーザに呈示する主題・内容キーワードの取捨選択方式の確立，Web文書での検索結果全体閲覧を可能にすることである．

参考文献

[1] 松倉 健志, 平田 陽一, 田中 克己: “話題構造に基づくWebページ間の意味的関係の発見,” DEWS,Mar,2001.

[2] Takeshi Matukura,Hiroyuki Kondo,Yoichi Hirata, Katsumi Tanaka: “Discovery of Semantic Relationships among Web Pages Based on Web Topic Structures,” DS-9,Apr,2001.

[3] 長尾 眞, 黒橋 禎夫, 白木 伸征: “単語の出現密度分布を用いた語の重要説明個所の特定,” 情報処理学会論文誌,Vol.38,No.4,pp.845-854,Apr,1997.