3U-3

コンテンツ・ベースド・フィルタリングを用いた Web 情報検索システムの提案

萩田 盾一† 辻 秀一‡
東海大学大学院工学研究科† 東海大学電子情報学科‡

1．はじめに

今日、ADSL や LDSL などの高速なデータ通信技術の登場や加入料の低価格化などによってインターネットに加入する人が大幅に増加し、それにともないホームページを開設する人も増加する傾向にある。そして現在のＷｅｂ情報量は日本だけでも１億ページを超えるまでになっており、インターネットは生活の一部を担う情報ツールとなっている。しかし、インターネットはその膨大な情報量から、そこから欲しい情報だけを探し出すといった作業はきわめて困難である。そこで、情報を探し出すための手段である「検索」に注目して、個人が必要とする情報だけを適切に探し出す個人支援を可能にした Web 情報検索システムを提案する。

2．従来技術

従来の情報フィルタリング技術をしてコンテンツ・ベースド・フィルタリングと呼ばれるものがある。この技術は正しいの要約に含まれる情報によりフィルタリングをかける技術であり、decision tree やnaive bayesやnearest neigh-borなどの技術によって構成されている。decision treeとはデータの属性の違いなどから分類していく方法で、naive bayes とは確率を用いた分類法でnearest neigh-borとは分類したいデータと正しく分類されているデータとの類似度を計算し、分類したいデータに似ている正しく分類されているデータの上位N個の分類を求める。そのN個のデータの分類を用いた投票を行い、カテゴリを決定するというものである。

そしてコンテンツ・ベースド技術を使用した従来技術としてモバイル情報提供サービスのgoopasが上げられる。このサービスは自動改札機の通過データを活用して利用者の場所・時間・行動をリアルタイムに把握し、あらかじめ設定された利用者の属性・嗜好を考慮して状況に応じた情報を提供する。これにより個人の要求にマッチした、その場所、その時間に適した情報を得ることができる。改札を通過した直後に電子メールで情報が提供される。[4]

The proposal of the Web information retrieval system using contents-besudo filtering.

†Junichi Hagita †Graduate School of Engineering, Tokai University
‡Hidekazu Tsuji ‡School of Information Technology and Electronics, Tokai University

3．提案方式

3-1 Web 情報検索システム全体構成

下記の図１は Web 情報検索の全体構成図である。このシステムはまず始めにインターネット上から気に入った Web ページを探し出すことから始まる。そして次にその Web ページから次回に使用できるような検索キーワードとなる単語を抽出し、検索キーワードデータベースに保存する。そして次回からの検索時には検索キーワードデータベースから必要な検索キーワードを取り出し、検索式を作成して検索をかけるというものである。

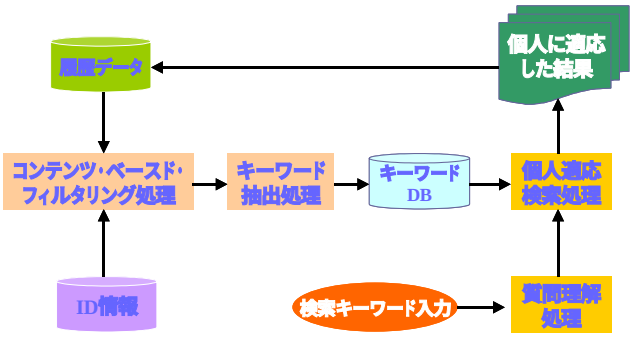

図１ 個人適応型 Web 情報検索全体構成図

3-2 検索キーワードを保存するまでの流れ

ここでは図２に示すように履歴データより検索キーワード抽出までの流れを説明する。まず履歴データを ID 情報を使用したコンテンツ・ベースド・フィルタリングにより、より個人が好む Web ページを限定する作業を行う。そして次にコンテンツ・ベースド・フィルタリング処理されたデータから検索キーワードに使用できる単語を抽出するキーワード抽出処理を行い単語（検索キーワード）を抽出しその単語をキーワードデータベースへ保存するというような流れである。

次に処理の具体例の説明をする。

① ここでは ID 情報にある程度の情報の好み、例えば映画の好みはアクションなどや音楽の好みはロックであるというような情報を元にコンテンツ・ベースド・フィルタリングを行う。

② ここでは処理された Web ページから単語を抽出する処理を行う。具体例をあげると、不要な HTML のタグをはずし、純粋なテキストデータにし、さらに形態素解析により、単語分割を行う。そ

して使われている頻度の高さによる重み付け、また形態素解析による品詞による重み付けを行い上位の単語を抽出するといった例を考えている。

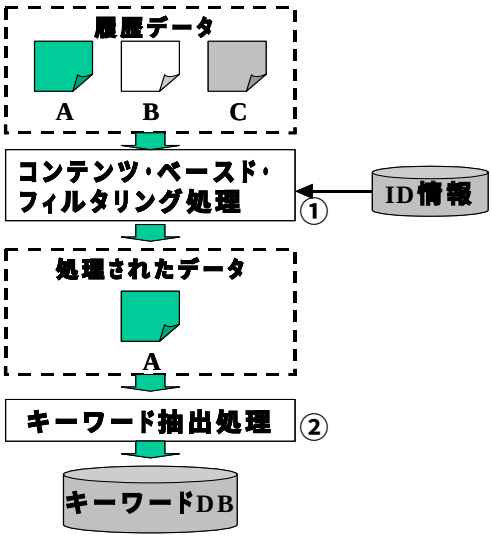

図２　キーワード保存までの処理フロー

3-3　キーワード DB を使用した検索の流れ

ここでは図３に示すようにキーワードデータベースを使用した検索を説明する。まず始めにユーザーは検索キーワードを入力する。そしてその検索キーワードを質問意味理解部に送りどのような検索をかけたいのかを理解し、次にその理解した質問意味によって図２で作成したキーワードデータベースから適切なキーワードを取り出し、検索式作成部によって新たな検索式を作成し検索をかけるというような流れである。

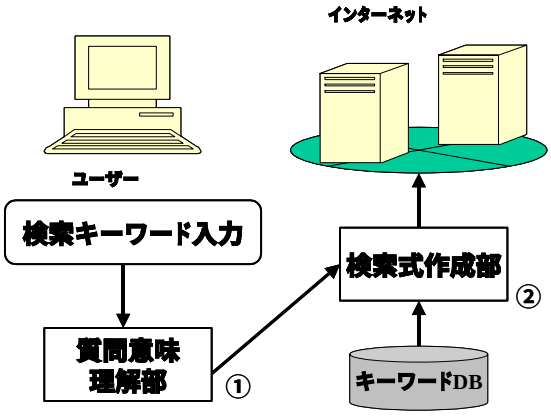

図３　キーワード DB を使用した検索の構成図

次に①、②の質問意味理解部と検索式作成部の具体例を上げる。まずユーザーの検索キーワード入力は検索対象の名詞一語だけの入力を考えている。これは出来るだけユーザーの手間を省きたいということを考えてのことである。

① 　この時の質問理解部では入力検索キーワードと辞書データベースとのマッチング処理を考えている。そして単語の意味にはその単語のカテゴリなどを単語の意味として次の処理に情報を送るという作業を行う。

② 　ここでは意味理解部により選択されたカテゴリによってキーワードデータベースより適切なキーワードを取り出し、新たな検索式を作成する。

４．考察

本提案方式の利点としてはまず第一にキーワードデータベースを使用した新たな検索式を作成により検索を絞り込むためのキーワードを入力する手間が省けるといった利点がある。そして第二に履歴データと ID 情報を使用することから個人の好みが反映された検索を行えるという利点があげられ、以上の利点により本提案は Web 情報検索において有効な手段であるといえる。

５．おわりに

本研究では、個人支援 Web 情報検索におけるのモデルの提案を行った。これからの Web 情報検索における「個人支援型 Web 情報検索」は、ますます情報量が増大していくであろうインターネットでの検索に非常に有効な手段であると思われる。

＜参考文献＞

[1]　高橋英史朗:「個人履歴情報を用いた Web 情報検索方式の提案」　情報処理学会第６５回全国大会／３U-3　、２００３年３月.

[2]　阿部秀尚:「構成型メタ学習と選択型メタ学習の比較評価」　人工知能学会第１７回全国大会／１F5-07、２００３年３月.

[3]　Stuart Russell･Peter Norving:エージェントアプローチ　人工知能

[4]　中尾寿朗,宮崎秀樹,藤本幸一,竹林一:「所･時間・行動を起点とした情報配信システム「goopas(グーパス)」－自動改札システムを利用したモバイル情報サービスの概要－　情報処理学会第６５回全国大会／６J-1　、２００３年３月.

[5]　Puntheeranurak Sutheera, Hidekazsu Tsuji:「Mining Web logs for Personalized Recommender System」 情報処理学会第67回全国大会／２E-4、２００５年３月.