ユーザに特化した情報収集エージェントの作成*

高砂　信吾
九州大学大学院システム情報科学府
知能システム学専攻†

長谷川　隆三, 藤田　博
九州大学大学院システム情報科学府
知能システム学部門　‡

1　はじめに

近年 WWW に蓄積される情報量は急速な勢いで増加を続けており，WWW から必要な情報を見つける為に多くのユーザが用いる手段として，既存の検索サイトを利用することが挙げられる．しかし大量の検索結果から必要な情報を選択する為には，ある程度の知識と経験が必要とされ，多くの初心者にとっては検索サイトを使いこなすことは容易ではない．WWW 上の情報検索におけるこのような問題を解決する１つの手段として，ユーザに特化した情報収集エージェントが Web ページを個人の計算機上に収集する手法が考えられる．収集されたページはユーザの興味を反映したページで構成される為，後にそのデータベースを検索に利用すれば，質の高い検索結果が期待できる．ここでユーザが興味を持つページを優先して収集する為に，情報収集エージェントに対しユーザの興味を与える必要がある．

本稿ではエージェントにユーザの興味を与える為に，エージェントが収集してきたページ集合に対して，ユーザが興味を持つページの評価を行い，そのページのリンク元のページに含まれるテキスト情報をエージェントに与える事で，エージェントの学習を行う．これによりエージェントは，ユーザが興味を持つページを優先して収集する．この手法ではまず，エージェントが WWW 上を探索している間に，収集されたページに対し任意時間においてユーザからの評価が行われる．そしてユーザによって興味があると評価されたページから，そのリンク元のページを求め訓練データとして使用する．最後に，エージェントが現在見ているページが，WWW 上の興味のあるページから見たどのリンク階層であるかを，訓練データから推定する．

以下の節では，上記の手法を用いた情報収集エージェント作成を行う．

2　WWW 上の探索

情報収集エージェントが，ページに含まれるリンクを辿りながら新たなページを取得する手法で一度探索したページを再び探索しないとすると，WWW 上のページとリンクは図１のようにループのない木構造でモデル化できる．ここで H のページが目的のページであるとすると，A，D，H の順にページを収集することが，最も効率の良い探索となる．本稿ではこのような探索経路を実現する為に，Web ページに含まれるテキスト情報を用いる．

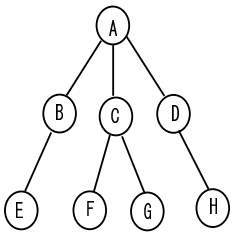

図 1: WWW Model.

この手法では，まず目的のページのリンク元のページのテキスト情報をエージェントの探索を効率化する為の訓練データとして用いる．訓練データは，目的のページからのリンク距離 j に応じて，クラス C_j に分類される．図１の例では，目的のページ H に対して D がクラス C_1 に，A がクラス C_2 に分類される．目的のページは，ユーザの評価によって定められる．ユーザはエージェントが収集してきたページに対し，興味があるか，もしくはないかの判定を行い，興味があるページを基にしてリンク元のページを求め訓練データを作成する．訓練データの各クラスに探索候補ページを分類することで，エージェントは目的のページに近いと思われるページから優先的に探索を行う．

3　テキスト分類

本稿では，Web ページから抽出したテキストの分類を単純ベイズ分類器 [3] を用いて，知識獲得及びページ分類を行う．さらにテキスト分類に必要とされる計

*Building Crawler for User-Specific Web Search Engines

†**Shingo Takasago**, Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University

‡**Ryuzo Hasegawa, Hiroshi Fujita**, Department of Intelligent Systems, Graduate School of Information Science and Electrical Engineering, Kyushu University

算量の増大を防ぐ為に，TF-IDF 法を用いて分類器内に存在するキーワードの刈り込みを行う．

3.1 キーワードの刈り込み

単純ベイズ分類器は訓練データに含まれる総語彙数に比例して計算量も増加するので，分類器の訓練データが蓄積される度に，エージェントの探索時間が増加する．そこで，分類器の更新を行う際に，TF-IDF（Term Frequency Inverse Document Frequency）法を用いて，分類器中に含まれる単語の刈り込みを行う．TF-IDF 値 $v(w)$ は以下の式で求められる．

$$v(w) = \frac{f^t(w)}{f^t{}_{max}} log \frac{N}{f(w)} \quad (1)$$

ここで，$f^t(w)$ はテキスト t に含まれる単語 w の数，$f^t{}_{max}$ はテキスト t 内の総単語数，N は総テキスト数，$f(w)$ は単語 w が含まれるテキスト数を表す．$v(w)$ がある閾値を満たさない単語はテキスト内から除去することにより，分類器の計算量増大を防ぐ．

3.2 ユーザによる Web ページの評価

エージェントが WWW 上を探索中に，エージェントが収集してきた Web ページに対し，そのページに興味があるかどうか任意時間でユーザが評価を行い，分類器の更新を行うことでエージェントの探索を効率化する．ユーザによる評価は，収集してきたページが興味のあるページとそれ以外のページに分類されることで行われる．最後に，興味があるページのクラスに分類されたページのリンク元のページを収集されたページ集合から求め，そのページ集合はそれぞれのリンク階層に応じてエージェント内で用いられる分類器の更新に用いられる訓練データとなる．

4 エージェントの探索システム

本稿で提案する情報収集エージェントを用いた探索システムの概要を図２に示す．エージェントは，収集してきたページ集合から，目的のページへ近いと思われるページ（キュー番号のより低いキューに収められた先頭のページ）から優先して抜きだし，ページ内に含まれるリンクの抽出を行う．得られたリンクから情報収集エージェントがリンク先のページを入手する．得られたページは HTML 解析を行いデータベースへ蓄積される．また得られたページは，そのページに含まれるテキストとベイズ分類器を用いてクラスの決定が行われ，各クラスに対応するキュー (C_j であれば $Queuej$) へ挿入される．

データベース内でまだ評価が行われていないページに対し，任意時間でユーザによる評価が行われる．それにより新たな訓練データが作成され，分類器の更新が行われる．

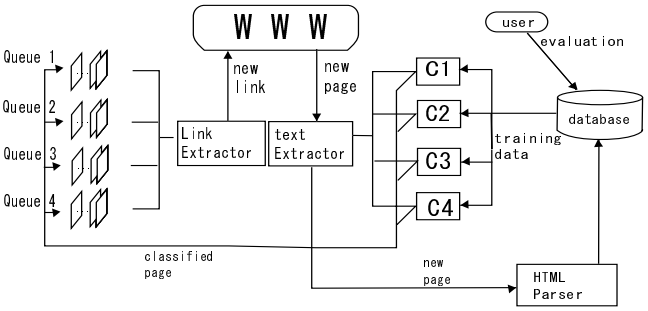

図 2: Overview of the system.

5 まとめ

本稿では，ユーザに特化した情報収集エージェントの作成の 1 つの手法を提案した．この手法ではまず，エージェントが収集してきたページ集合に対してユーザが評価を行い，分類器の更新を行う．そして，エージェントは分類器を基にして探索を行うページの決定を行う．その結果エージェントが，ユーザが興味を持つページを効率よく収集することを目的としたものである．この手法を用いて，WWW 上からユーザの興味を反映した情報を優先して収集することにより，ユーザに特化した検索エンジンのデータベースを作成できると期待できる．また今後の課題として，提案システムを実装し評価実験を行うことや，ユーザによる評価法の新たな手法について検討を行うことなどが挙げられる．

参考文献

[1] A. McCallum,K. Nigam, J. Rennie, and K. Seymore :“A Machine Learning Approach to Building Domain-Specific Search Engines”,*IJCAI-99*, 662-667, 1999.

[2] M. Diligenti, F.M. Coetzee, S. Lawrence, C.L. Giles and M. Gori :“Focuced Crawling Using Context Graphs”, *VLDB-2000*, 527-534, 2000.

[3] Mitchell, T. M. “*Machine Learning*”, McGrawHill, 1997.

[4] 松本裕治，北内啓，山下達雄，平野善隆，松田寛，高岡一馬，浅原正孝:“形態素解析システム「茶筌」version 2.3.3　使用説明書”，奈良先端科学技術大学院大学情報科学研究科自然言語処理学講座，http://chasen.aist-nara.ac.jp/，2003.