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 : 修士課程

長谷川, 隆三
九州大学大学院システム情報科学研究院知能システム学部門

藤田, 博
九州大学大学院システム情報科学研究院知能システム学部門

越村, 三幸
九州大学大学院システム情報科学研究院知能システム学部門

https://doi.org/10.15017/1515922

出版情報：九州大学大学院システム情報科学紀要. 9 (1), pp.25-29, 2004-03-26. 九州大学大学院システム情報科学研究院
バージョン：
権利関係：

九州大学大学院
システム情報科学紀要
第9巻 第1号 平成16年3月

Research Reports on Information Science and
Electrical Engineering of Kyushu University
Vol.9, No.1, March 2004

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* ・ 長谷川 隆三** ・ 藤田 博** ・ 越村 三幸**

Building Crawler for User-specific Web Search Engines

Shingo TAKASAGO, Ryuzo HASEGAWA, Hiroshi FUJITA and Miyuki KOSHIMURA

(Received December 12, 2003)

Abstract: Today, the volume of available data on the WWW becomes very huge, and searching information from the WWW is a difficult task for a novice user even if he/she uses the standard search engines. One solution to the problem is to build a user-specific search engine, the database of which includes a large number of web documents required for a user. In this paper, we present a method of building a crawler aiming to search the subset of the WWW related to on-topic pages. We show an effective strategy for leading the crawler to on-topic pages by using naive Bayes text classifier trained by an evaluation of pages gathered by the crawler.

Keywords: Crawler, Naive Bayes, Text classification, TF-IDF, World Wide Web

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WWW上の情報を個人の計算機上に収集する手法が考えられる．個人の計算機に収集されたWWW上の情報は，ユーザに特化した検索エンジンのデータベースとして使用される．このデータベースは，データベースに含まれるページ全体に対しユーザの興味を反映したページが多く含まれる為，既存の検索サイトを用いるよりもより質の高い検索結果が期待できる．

ユーザに特化した情報収集エージェントの性能を決める重要な点の1つとして、WWW上からユーザが求める情報を効率よく収集することが挙げられる．ここでの効率の良い収集は，エージェントがWWW上でリンクを辿りながらページ集合を収集する際に，ユーザが必要としないページを避けながら，必要とするページを多く収集できる探索を行うことで達成される．その為には，エージェントが収集した情報に対しその内容に応じた報酬を適切な手法で与えることでエージェントの学習を行う方法が考えられる．

エージェントがWeb上から特定のページを収集する手法として，コンピュータサイエンスの研究論文専門の検索エンジンであるCora[1]で用いられる手法が挙げられる．ここで用いられるエージェントは研究論文のページを収集することを目的としている．エージェントは目的のページを見つけると，以前の探索で得られたリンクやテキスト情報から学習を行うことで，探索を徐々に効率化させる手法を用いている．またDiligentiらは，研究論文に関するページを目的のページとし，既存の検索エンジンを利用して目的のページへのリンクを持つページ集合を求め，そのページのテキスト情報から学習を行っている[2]．

本稿では，ユーザが興味を持つページを収集することが目的となる．そこで，ユーザ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，さらにそ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に与える事で，エージェントの学習を行い探索を効率化させる手法を提案する．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し探索の効率化を行う．

本稿の構成は以下のとおりである．2章でWWW上での情報収集エージェントの探索効率化について述べる．3章ではテキスト分類手法の1つである単純ベイズ分類器

平成15年12月12日受付
* 知能システム学専攻修士課程
** 知能システム学部門

(Naive Bayes Classifier)を取り上げその分類器の作成や訓練データの更新などについて説明する．4章でエージェントが収集してきたWebページ集合に対するユーザの評価法を紹介し，5章でエージェントが行う探索システムの説明を行う．6，7章では本稿で提示したエージェントに対する評価実験を行い，最後に8章で本稿のまとめを行う．

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**Fig 1**のようにループのない木構造でモデル化できる．エージェントが現在取得したページをAとすると，

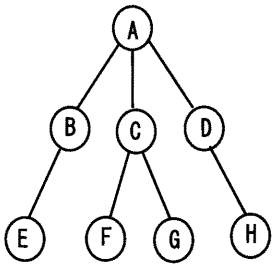

Fig. 1　WWW Model.

次に取得可能なページは，B，C，Dとなる．ここでエージェントがBのページを先に取得すると，次に取得可能なページはC，D，Eとなる．Hが目的のページ（ユーザが興味を持つページ）であるとすると，エージェントが，ページHが目的のページであるとAのページから推測することにより，A，D，H，の順にページの取得を行うことが最も効率の良い探索となる．

実際にユーザが現在見ているページから目的のページを探索する時，現在のページの内容から，ページに含まれるリンク先のページの内容を予測して，リンクを辿りながら目的のページを発見する手法が考えられる．これをエージェントに当てはめると，エージェントがWWW上を探索中にページの解析を行い，目的のページにより近いリンクを選択する知識をページに含まれるテキスト情報から取得できれば，効率的な探索が可能である．

本稿で用いる手法は，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．**Fig 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．与えられた訓練データを参考にして，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ことで，効率の良い探索を図る．

3. Webページのテキストの分類

Webページ取得の為に，リンクを辿りながらWWW内を探索する手法を考えると，最も単純な戦略は幅優先探索である．この探索法はWeb上全てのページを収集することを目的とする為，ユーザに特化した検索エンジンを構築しようとするには不向きである．よって本稿ではこの手法は採用しない．しかしこの手法は各ページに対する知識の獲得と目的のページに近いページの推定を行う必要がない為，単ページに対する実行時間からみたパフォーマンスは最良である．Webページからの知識獲得やページの分類には多くの計算量を必要としない手法で行われるべきである．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 単純ベイズ分類器

単純ベイズ分類器は，いくつかにクラス分けされたテキスト集合Tが訓練データとして与えられた時，新たに与えられたテキスト$\hat{t}$がどのクラスに属するかを，訓練データを用いた最大事後確率の計算により求める．

テキスト$\hat{t}$は，テキストに含まれる単語$w_{\hat{t}k}$の連言とみなす．単純ベイズ法では，あるテキスト$\hat{t} = (w_{\hat{t}1}, w_{\hat{t}2}, \cdots, w_{\hat{t}k}, \cdots)$がクラス$c_j$に属する確率$P(c_j \mid \hat{t})$を以下の式で求める．

$$P(c_j \mid \hat{t}) \propto P(c_j)P(\hat{t} \mid c_j)$$

$$\propto P(c_j) \prod_{k=1}^{|\hat{t}|} P(w_{\hat{t}k} \mid c_j) \tag{1}$$

$P(c_j)$は，任意のテキストがクラスc_jである確率，$P(\hat{t} \mid c_j)$はクラスc_j内にテキスト$\hat{t}$が存在する確率である．ここで訓練データから、$P(c_j), P(w_{\hat{t}k} \mid c_j)$を$P(c_j), P(w_{\hat{t}k} \mid c_j) \neq 0$とする為の補正値を加えて求める．

$$P(w_{\hat{t}k} \mid c_j) = \frac{1 + n_{\hat{t}k}}{\mid V \mid + n} \tag{2}$$

n_{ik} : c_jの訓練テキストに$w_{\hat{t}k}$が出現した回数
n : c_jの訓練テキストに出現した単語数
$|V|$: 全訓練テキストに出現した語彙の数

$$P(c_j) = \frac{1 + t_{c_j}}{|C| + |T|} \tag{3}$$

t_{c_j} : c_jに分類された訓練テキスト数
$|T|$: 総訓練テキスト数
$|C|$: 総クラス数

新たなページに含まれるテキスト$\hat{t}$に対し$\mathrm{argmax}_{c_j \in C} P(c_j \mid \hat{t}_i)$を求めることでクラス分類を行う．

3.2 テキストからのキーワード抽出

テキストの解析は，形態素解析システム「茶筅」[5]を利用して行われる．また，テキスト内には，テキストを特定するキーワードになり得ない品詞を持つ単語が含まれる為，分類器内で用いる単語をその品詞により制限する．本稿で採用する品詞は以下に示される．

・形容詞　自立，接尾，非自立
・動詞　自立
・名詞　サ変接続，一般，形容動詞語幹，固有名詞

また語幹が同じ単語は全て基本形に改める．

3.3 分類器の更新

エージェントの探索中の任意時間に，エージェントが収集してきたページに対しユーザが評価を行う．ユーザが評価を行ったページのリンク元のページを，収集してきたページ集合の中から求め，ページに含まれるテキスト情報によって分類器の訓練データが蓄積される．

しかし，単純ベイズ分類器は訓練データに含まれる総語彙数$|V|$に比例して計算量も増加するので，目的のページを発見する度に分類器の訓練データが蓄積されると，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法は，テキスト集合の各単語に対して重み付けを行う．TFは，テキスト中で繰り返し生起する単語はそのテキストにおいて重要な概念であることを重み付けに用いる．しかし，多くのテキストに生起する単語は，テキストを特定する性質を持たず，キーワードとして適していない．そこで、ある単語がどのくらい特定性を持つかを，IDFを用いることで重み付けに反映させる．訓練データ内の単語wのTF-IDF値$v(w)$は以下の式で求められる．

$$v(w) = \frac{f^t(w)}{{f^t}_{max}} log \frac{N}{f(w)} \tag{4}$$

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．分類器の更新は以下の手順で行われる．

1. 分類器内のテキスト集合Tに新たにユーザによりクラス分けされたテキスト集合$\hat{T}$を追加．
2. テキスト集合T内の全てのテキストtでTF-IDF法を用いた単語の刈り込みを行う．

4. 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(on-topic page)とそれ以外のページ(off-topic page)に分類されることで行われる．しかし，収集してきた全てのページに対してユーザが評価を行うことは難しい為，まず収集されたページからランダムにn個のページを選び，ユーザが興味のあるページとそれ以外のページに分類する．

さらに，ユーザによって評価が行われ2つのクラス(C_{on}とC_{off})に分類されたページ集合を訓練データとし，残りのページ集合に対してもページに含まれるテキストから分類が行われる．この分類は単純ベイズ法を用いて行われる．

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．

新たに収集されたWebページ集合Dに対しユーザによる評価を行い，エージェント内で用いられる分類器に追加されるテキスト集合$\hat{T}$となるまでの手順を以下に示す．

1. 収集されたWebページ集合Dのランダムに選ばれたページn個の集合D_tがユーザによりC_{on}, C_{off}分類される．
2. 残ったページ集合$D - D_t$が，D_tを訓練データとしてC_{on},C_{off}に分類される．

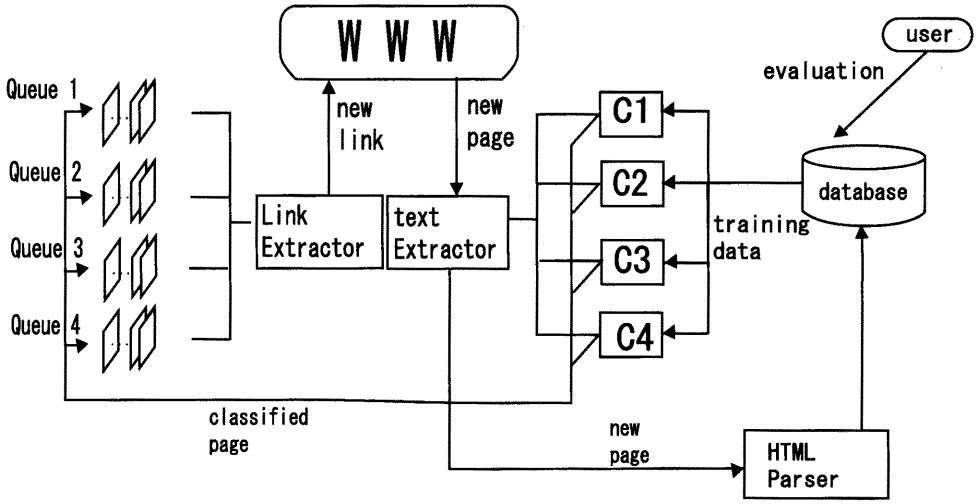

Fig. 2　Overview of the proposed method.

3. C_{on}内の各ページに対し，リンク元のページをデータベースから求め，テキスト集合$\hat{T}$を作成する．

5. 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**Fig 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解析された後，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た後に，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．

6. 評　価　実　験

提案手法の有効性を実証する為に，収集された興味のあるWebページの数について一般的な探索法と比較する．ここで一般的な探索とは，網羅的な幅優先探索を意味し，ユーザによる評価を探索に用いない手法である．幅優先探索の手順は，以下のようになる．

1. 始点となるページP_0に含まれるリンク集合を，キューQに追加．
2. Qの先頭のリンクをQより取り出す．
3. 取り出したリンクに対応するWebページPを取得
4. Pに含まれるリンク集合をQに追加．
5. 2.に戻る．

また，提案手法を用いた探索ではキーワードとして英単語を用いない為，ページ内にキーワードが1つも含まれないような英語表記のページについては，探索は行わないという制限を加える．

実験は九州大学内のWebページ集合を用いる．今回探索を行う範囲として，九州大学大学院システム情報科学研究院のページ(http://www.isee.kyushu-u.ac.jp/)を始点として，幅優先探索により得られた5,159個のページを用いる．ページ内に含まれるリンクの総数は34,581個である．また，ユーザが興味を持つページとして，各研究室のトップページと最初から定めておく．今回用いるWebページ集合の中には190個の目標ページが含まれる．

このページ集合内で探索を行い，収集された目標ページ数を比較した．探索結果を比較する為に本稿では3つのエージェントを用いた．

- Breadth-First Crawler(BFC):幅優先探索法を用いた探索を行う．
- Crawler-500(C500):提案手法を用いた探索を行う．新たなページが500収集される度に分類器の更新が行われる．
- Crawler-200(C200):同じく提案手法を用いる．新たなページが200ページ収集される度に更新を行う．

提案手法で用いられる分類器のクラス数はいずれも3とする．また，ユーザによる評価は，ある一定数のページの収集が行われた時に，収集されたページ集合の中からランダムに30個のページを抽出し，そのページ集合内に含まれる目標ページを分類器の更新に用いることで行われる．

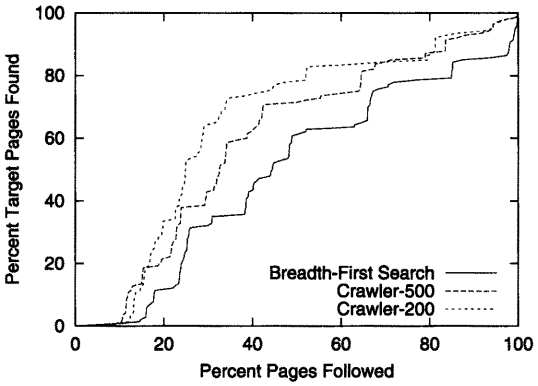

Fig. 3 Crawling Kyushu University Data Set.

Table 1 Predictive words for Crawler-200.

class 1	class 2	class 3
研究(380)	工学(442)	情報(646)
工学(341)	システム(200)	専攻(612)
専攻(118)	研究(188)	システム(475)
科学(99)	講座(90)	科学(283)
環境(84)	材料(82)	電気(273)
ページ(83)	科学(81)	事務(215)
大学(79)	教授(74)	平成(179)
情報(77)	エネルギー(68)	入学(115)
大学院(69)	電気(66)	工学(110)
材料(66)	大学(58)	環境(109)

7. 結果と考察

Fig 3が実験結果をグラフにまとめたものである．探索初期において，ユーザによる評価が行われていない段階では，分類器内に訓練データが存在しない為，各エージェントは同様の探索を行っている． ユーザによる評価が行われると，提案手法を用いたエージェントとBFCの間に違いが生じるのが分かる．ページ全体の約35％を収集した時には，BFCに対してC500は約1.58倍，C200は約1.94倍の目的ページを発見している．このように，探索途中により多くの目的のページを発見できていることから，実際のWeb上での探索において，より多くの目的のページを収集できると考えることができる．データベースの作成において，この探索手法を用いたページ収集により，そのデータベース内にユーザが興味を持つページをより多く収集することが可能になる為，データベースを検索に利用する際に，より質のよい検索結果が期待できると考えられる．またC500とC200の比較から，ユーザによる評価が頻繁に行われれば，エージェントはより効率的な探索を行うことが可能であることが分かった．

Table 1は，ページ全体の約35％(1800ページ)を収集した際に，C200の分類器内に含まれる単語をクラス別に並べたものである．ただし比較の為，TF-IDF法により削除された単語についても表示している．()内の数値は，分類器内の単語数を表している．上位のクラスには“研究”という研究室を示す単語，下位のクラスには“平成”，“事務”などの研究室のページとは関係のない単語が多く含まれることが確認できる．このような単語の振り分けが行われている為，分類器が効果的に作用していることが分かる．

以上の実験結果から，ページの評価を探索に反映させる為に，ページに含まれるテキスト情報を利用して探索経路を決定する事で，目的のページをより多く発見できることが確認できた．

8. 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，ユーザが興味を持つページをエージェントが効率よく収集することを目的としたものである．評価実験の結果，収集された目的のページを訓練テキストとしてエージェントに与える事で，探索の効率化が行われていることが分かった．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，ユーザによる評価を行う際のインターフェイスを作成し，被験者による評価実験を行うことが挙げられる．

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) 嶋津恵子，山根洋平，門馬淳仁，桜井哲志，古川康一，“テキストデータの内容に基づく相関ルールのクラスタリング実験”，第50回人工知能基礎論研究会資料，55-62，2002.
5) 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「茶筅」 version 2.3.3 使用説明書”，奈良先端科学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.