ブックマークエージェント: WWW における協調的情報フィルタリング

Bookmark-agent: Information Sharing of WWW

○ 森 幹彦　　山田 誠二
Mikihiko Mori　　Seiji Yamada

東京工業大学 大学院 総合理工学研究科
IGSSE, Tokyo Institute of Technology

We propse a bookmark-agent system that shares information in bookmark files. The agent requests for other agents to search their own bookmarks. Since the acquired information is filtered beforehand by users, the Bookmark-agent is able to output more precise Web pages than a search engine. We finally made experiments by six users, and found out the Bookmark-agent is a promising approach to share URLs in a small community.

1 はじめに

インターネットの急速な広がりとともに，ユーザは World Wide Web (WWW) の文書情報を検索することに多大な労力を払わなければならなくなってきている．そのため，検索エンジンと呼ばれる，キーワードを与えることで Web ページのを検索ができるシステムの開発が進められている．しかし，これらの検索エンジンは必要以上の検索結果を提示してしまうのが現状である [3][6].

そこで我々はブックマークエージェントと呼ばれるシステムを提案した [4][5]．すなわちブックマークエージェントと呼ばれるプログラムは，ユーザのブックマーク*を参照し，さらに他のエージェントとの通信により，他のユーザのブックマークの URL をも参照して，現在ユーザの見ている Web ページと類似したページを提示する．このように，ブックマークエージェントは，ユーザの明示的要求を必要とせずに，ユーザの欲する URL 情報を提示できる．また，ユーザは，ブックマークエージェントに，明示的にキーワードを提示することも可能である．

2 ブックマークエージェント

2.1 基本機能

ブックマークエージェントは，1 ユーザあたりに 1 つ起動される．つまり，ブックマークエージェントは，一人のユーザに対し情報検索支援を行う．ブックマークエージェントの機能を以下に挙げる．

- **新規 URL のキーワード獲得：**ユーザのブックマークを監視し，新たな URL の追加があった場合，その URL の HTML ファイルからキーワードを抽出してキーワードデータベースに追加する．
- **類似ページの提示：**ユーザが現在見ている Web ページからキーワードを抽出し，それを元にそのページの類似ページの URL をユーザに提示する．さらに，ユーザが明示的にキーワードをブックマークエージェントに与えて，類似ページを提示させることも可能である．ここで，**類似ページ**とは，後述する類似度があるしきい値以上のページを意味する．類似ページの検索方法は，以下の 2 通りがある．
  - ブックマークエージェントが担当しているユーザの URL データベースを検索する．
  - 他のブックマークエージェントに URL 情報検索を依頼する．
- 他のブックマークエージェントから URL 情報検索の依頼があった場合，担当ユーザの URL データベースで検索を行い，類似ページを返す．

2.2 キーワード抽出と類似度

テキストからのキーワード抽出は，出現頻度を用いるものを始めとして，さまざまな方法が提案されている [1] が，HTML という構造化された文書を扱うことから，本研究では，以下に述べるようなタグ構造に注目したキーワード抽出を用いている．

HTML ファイルにおいて，下記のタグ中に含まれる単語（名詞）に対し，括弧内の重みを加算していき，その重みの上位 5 つをその URL のキーワードとする．

・<META>(10) ・<TITLE>(10) ・<Hn>(6, 4) ・<EM>(1)

また，ページ間の類似度はそれぞれのキーワードの積集合の要素数とする．よって，類似度は 1 から 5 の値をとる．

3 実験

本節では，ブックマークエージェントの性能評価のための実験を行う．

ここで，ブックマークエージェントの評価基準として適合率を挙げる．適合率の定義を以降に示す．

連絡先: 森幹彦 東京工業大学大学院総合理工学研究科
〒 226 横浜市緑区長津田町 4259 番地 TEL,FAX: 045-924-5218
email: mori@ymd.dis.titech.ac.jp

*一度見た興味ある Web ページを記録するファイルのことである．

3.1 適合率の定義

情報検索の効率の定量的評価の方法として，呼出率と適合率がある [2]．これらの定義は，以下の通りである．

- **呼出率：** 質問と適合する文献総数に対する検索された適合文献の割合．
- **適合率：** 検索された文献総数に対するその中での適合文献数の割合．

ここで，検索の母集合となる WWW の文書数を計ることは現実的に不可能なため，呼出率を求めることはあきらめる．一方，適合率は検索された文書数を母集合とし，その中で正しく検索された文書数の率である．したがって，適合率を検索効率の指標とすることにする．

検索エンジンの検索結果では時として数千から数万件の URL を提示することがあり，この場合，適合率を調査することすら困難な状態となってしまう．しかし，検索エンジンでは URL を並び替え，与えられたキーワードの内容により近いと思われる URL を先に表示する機能がある．したがって，上位 20 位までに入った URL の適合率を求めることで近似的にそのキーワードに対する適合率とする．

3.2 実験に参加するユーザ

ブックマークエージェントの評価実験には，表 1に示す 6 名が参加する．各ユーザはそれぞれブックマークから生成されたキーワードデータベースを所有する．実験中は，提示された URL がそれがどのユーザから与えられたものなのかを識別することができる．

注目することには，このユーザグループではみなフリー OS である Linux を使用するユーザであり，AI，エージェント，ロボットを研究しているということである．

3.3 実験結果

ブックマークエージェントには，各ユーザのキーワードデータベースを総合してその中に存在する任意のキーワードを 1 から 3 個ずつ与える．本実験では，ユーザ C が被験者になり，提示された URL が適合であるかどうかの判定を行った．また実験では，ブックマークエージェントと比較するために検索エンジン goo を用いた．

表 2，3，4はそれぞれ 1 から 3 個のキーワードを与えたときの適合率をまとめたものである．

4 考察

本節では，先述した実験適合率と貢献率の実験結果について考察する．ブックマークエージェントと検索エンジン goo では振る舞いが異なることがわかる．そして，ブックマークエージェントの特徴を発見した．

4.1 ブックマークエージェントの適合率

適合率は，検索するキーワードの単語としての性質によって特徴がある．

抽象的なキーワードは具体的な内容のキーワードよりも適合率が低い．例えば，キーワード ‘internet’ と ‘web’ は必ずしも特定の 1 つのものを表すだけではないく，場

表 1　ユーザの有するブックマーク中の URL 数

ユーザ名	A	B	C	D	E	F	合計
URL 数	10	356	137	85	167	71	782

表 2　キーワード 1 個を与えた場合の適合率

キーワード	適合率 (%)	
	ブックマークエージェント	検索エンジン
agent	100.0	5.0
robot	100.0	30.0
intelligence	100.0	10.0
ai	100.0	45.0
research	75.0	20.0
internet	26.7	15.0
web	35.7	10.0
linux	100.0	70.0
unix	100.0	40.0
science	33.3	50.0
software	91.7	45.0
faq	100.0	50.0
perl	75.0	75.0
java	100.0	45.0
sendmail	100.0	65.0
search	84.6	55.0
mac	100.0	60.0
macintosh	100.0	85.0
weather	100.0	75.0
dictionary	100.0	35.0
japanese	21.1	40.0
english	37.5	5.0
ml	75.0	35.0
archie	100.0	55.0
平均	81.5	44.2

表 3　キーワード 2 個を与えた場合の適合率

キーワード		適合率 (%)	
		ブックマークエージェント	検索エンジン
software	agent	100.0	30.0
web	agent	100.0	15.0
linux	software	100.0	35.0
free	software	100.0	10.0
english	dictionary	100.0	70.0
java	software	100.0	30.0
java	internet	100.0	15.0
web	robot	100.0	55.0
artificial	intelligence	100.0	45.0
linux	application	100.0	20.0
tokyo	university	100.0	50.0
mailing	list	100.0	75.0
fj	news	100.0	60.0
sports	soccer	100.0	35.0
平均		100.0	38.9

表 4　キーワード 3 個を与えた場合の適合率

キーワード			適合率 (%)	
			ブックマークエージェント	検索エンジン
ai	artificial	intelligence	100.0	45.0
agent	robot	autonomous	100.0	85.0
linux	tool	application	100.0	40.0
ai	lab	mit	100.0	90.0
english	japanese	dictionary	100.0	35.0
mailing	list	japan	100.0	70.0
tokyo	institute	technology	100.0	75.0
mac	news	information	100.0	30.0
平均			100.0	58.8

合によっては抽象的な意味を持つ．すなわち，‘internet’ は “Internet Program” という使い方で，“インターネットのための” という意味や “インターネット中の” という意味になる．また，‘web’ は “Someone’s Web” という使い方で，“だれそれのウェブサイト” という意味になる．もし，‘internet’ という単語を “インターネットの構造” や “インターネットの特徴” と言う意味でキーワードとして使えば，適合率は低くなる．例えば，表 2から，固有名詞の ‘linux’, ‘faq’, ‘mac(intosh)’ の適合率が 100%なのに対して，‘internet’ が 27.6%，‘web’ が 35.7%である．

キーワードが形容詞になりやすい単語の場合，その適合率は低い．例えば，‘web’ や ‘english’ は名詞にもなるが，後ろに名詞をとることで “Web Agent” や “English Dictionary” のように形容詞としても働く．このような単語の場合，単独での適合率よりも他の単語をとって熟語を作ったときの方が適合率は高い．すなわち，表 2と 3より，‘web, agent’ や ‘english, dictionary’ の適合率は 100%なのに対して ‘web’ が 35.7%，‘english’ が 37.5%となってしまう．

4.2　適合率の検索エンジンとの違い

グループ内で興味を共有するキーワードの場合，ブックマークエージェントは，興味を共有しないキーワードの時に比べて適合率が高くなる．例えば，‘agent’ はふつう，‘代理人’ や ‘代行’ の意味である．しかし，AI についてのグループであれば，その意味はコンピュータを使ったエージェントシステムという意味になるはずである．

表 2より，goo において，キーワード ‘agent’ の適合率は平均適合率の半分以下になっているのに対して，ブックマークエージェントではほとんど変わらないことがわかる．これは，ブックマークエージェントが，グループのメンバによって，あらかじめ AI 関連の興味によるフィルタリングされているためであると考えられる．

各キーワードに対する適合率において，ブックマークエージェントと検索エンジン goo に有意な差が見られるかどうかを調べるために，順位和検定を行った．キーワード数が 1 個，2 個，3 個のときに，エージェントと検索エンジンとの適合率を標本として検定を行った．ここで，それぞれの個数のときの検定統計量を T_1, T_2, T_3 とする．すると，

$$T_1 = 3.958973 > 1.96$$

$$T_2 = 3.997448 > 1.96$$

$$T_3 = 2.940588 > 1.96$$

となり，それぞれ有意水準 5%で有意な差がある．

この検定により，ブックマークエージェントは検索エンジンにはない URL 提示法を用いていることが明らかとなり，協調的情報フィルタリングが有効であることが示唆された．

4.3　興味の共有の有無による違い

まず，ユーザ同士が興味を共有している場合を考える．本実験におけるユーザは一般的な科学には興味がないが，AI，ロボット，エージェントシステムには興味がある．したがって表 2より，ブックマークエージェントにおいて，‘ai’, ‘robot’, ‘agent’ といった単語の適合率は ‘science’ という単語の適合率よりも高い．

一般的に，UNIX という単語は Linux という単語よりも有名であるので，提示 URL 数は UNIX のほうが多いはずである．しかし，ここで思い出さなければならないことは，本実験のユーザは皆 Linux ユーザであるということだ．したがって，キーワード ‘linux’ は ‘unix’ よりも多数の URL を提示している．

以上のことより，ブックマークエージェントはユーザの興味に影響を受けて URL 提示を行っていることがわかり，協調的情報フィルタリングがかかっていることを示している．

5　おわりに

本研究では，URL 情報であるブックマークを用い，情報検索を支援するブックマークエージェントを提案した．また，6 人のユーザによる実験を行った．ブックマークは，ユーザによってあらかじめフィルタリングされた URL 情報であると考えられる．したがって，検索エンジンがユーザが望まない Web ページを多数提示してしまうのと異なり，ブックマークエージェントは適度にフィルタリングされ，少数であるがユーザの要求に見合った URL 情報を，協調的情報フィルタリングによって提供することができた．

参考文献

[1] Salton, G. and McGill, M. J. : Introduction to modern information retrieval, *McGraw-Hill* (1983)

[2] 伊藤 : 情報検索, 昭晃堂 (1986)

[3] 武田 : ネットワークを利用した知的情報統合, 人工知能学会誌, Vol. 11 No.5, pp. 680-688 (1996)

[4] 森, 山田 : ブックマークエージェントによる WWW の URL 情報の共有, 第 11 回人工知能学会全国大会論文集, pp. 486–487 (1997)

[5] 森, 山田 : ブックマークエージェントによる URL の協調的情報フィルタリング, 人工知能学会第 29 回人工知能基礎論研究会論文集, pp. 7–12 (1997)

[6] 森田, 速水 : 情報フィルタリングシステム, 情報処理, Vol. 37 No.8, pp. 751–757 (1996)