IMPROVED REINFORCEMENT-BASED PROFILE LEARNING FOR DOCUMENT FILTERING

YAHYA MOHAMMED ALMURTADHA

FSKTM 2007 13
IMPROVED REINFORCEMENT-BASED PROFILE LEARNING FOR DOCUMENT FILTERING

By

YAHYA MOHAMMED ALMURTADHA

Thesis Submitted to the School of Graduate Studies, Universiti Putra Malaysia, in Fulfilment of the Requirements for the Degree of Master of Science

June 2007
Dedicated to

my parents, my wife,

my brothers, and my sisters
Abstract of thesis presented to the Senate of Universiti Putra Malaysia in fulfillment of the requirement for the degree of Master of Science

IMPROVED REINFORCEMENT-BASED PROFILE LEARNING FOR DOCUMENT FILTERING

By

YAHYA MOHAMMED ALMURTADHA

June 2007

Chairman : Associate Professor Hj. Md. Nasir Sulaiman, PhD
Faculty : Computer Science and Information Technology

Today the amount of accessible information is overwhelming. A personalized information filtering system must be able to tailor to current interests of the user and to adapt as they change over time. This system has to monitor a stream of incoming documents to learn the user’s information requirements, which is the user profile.

The research has proposed a content-based personal information system learns the user’s preferences by analyzing the document contents and building a user profile. This system is called RePLS; an agent-based Reinforcement Profile Learning System with adaptive information filtering. The research focuses on an improved terms weighting to measure the importance of the terms represent each profile called “purity term weighting”. The top selected terms are then used to filter the incoming documents to the learned user profiles. The agent approach is used because of its autonomous and adaptive capabilities to perform the filtering.
The proposed method was evaluated and compared with three Information Filtering methods, namely Rocchio, Okapi/BSS Basic Search System and Reinf, the incremental profile learning method. Based on the proposed method, a profile learning system is developed using Microsoft VC++ connected to Microsoft Access database through an ODBC. AFC kit is used to implement the proposed agents under RETSINA architecture. The experiments are carried out on the TREC 2002 Filtering Track dataset provided by the National Institute of Standards and Technology (NIST).

This research has proven that RePLS is able to filter the stream of incoming documents according to the user interests (profiles) learned by the proposed Purity term weighting method. Based on the experiments results, Purity weighting shows better terms weighting and profile learning than the other methods. The outcome of a considerably good accuracy is mainly due to the right weighting of the profile’s terms during the learning phase.

This research opens a wide range of future works to be considered, including the investigation of the dependency between the selected terms for each profile, investigating the quality of the method on different datasets, and finally, the possibility to apply the proposed method in other area like the recommendation systems.
Abstrak tesis yang dikemukakan kepada Senat Universiti Putra Malaysia sebagai memenuhi keperluan untuk ijazah Master Sains

PENAMBAHBAIKAN PEMBELAJARAN BERASASKAN PENGUKUHAN PROFIL BAGI PENYARINGAN DOKUMEN

Oleh

YAHYA MOHAMMED ALMURTADHA

Jun 2007

Pengerusi : Profesor Madya Hj. Md. Nasir Sulaiman, PhD

Fakulti : Sains Komputer dan Teknologi Maklumat

Dewasa ini jumlah maklumat yang sedia diperoleh adalah sangat memberangsangkan. Sesebuah sistem penyaringan maklumat peribadi mestilah berupaya disesuaikan dengan kehendak individu pengguna, serta mampu berubah sejajar mengikut perubahan masa. Sistem tersebut perlu mengawasi aliran masuk dokumen-dokumen pengguna bagi mempelajari keperluan maklumat mereka, iaitu profil pengguna.

Kajian ini mencadangkan sebuah sistem maklumat peribadi berasaskan kandungan, yang berupaya mempelajari kecenderungan pengguna melalui analisa kandungan dokumen beserta profil pengguna. Sistem ini dinamakan RePLS, iaitu sistem pembelajaran profil berasaskan pengukuhan melalui agen. Kajian ini tertumpu ke arah penambahbaikan pemberat terma yang digunakan bagi mengukur kepentingan sesuatu terma yang wujud dalam setiap profil. Terma-
terma terbaik yang terpilih akan digunakan bagi menyaring kemasukan dokumen dalam proses mempelajari profil pengguna. Pendekatan berasaskan agen digunakan kerana sifatnya yang autonomi dan boleh diadaptasi dalam menjalankan proses penyaringan.

Kaedah yang dicadangkan ini telah dinilai dan dibandingkan dengan tiga buah algoritma penyaringan maklumat yang lain, iaitu Rocchio’s Algorithm, Okapi/BSS Basic Search System dan Reinf, sebuah kaedah pembelajaran profil berperingkat. Berdasarkan kaedah yang dicadangkan ini, sebuah sistem pembelajaran profil telah dibangunkan dengan menggunakan Microsoft VC++ dan dihubungkan kepada pangkalan data Microsoft Access melalui ODBC. Kit AFC digunakan bagi membangunkan agen yang tersebut di bawah kerangka RETSINA. Eksperimen yang dibuat adalah menggunakan set data TREC 2002 Filtering Track oleh National Institute of Standards and Technology (NIST).

Kajian ini telah membuktikan bahawa RePLS berupaya menyaring ses sebuah aliran masuk dokumen mengikut kehendak atau profil pengguna. Berdasarkan keputusan eksperimen, RePLS menunjukkan prestasi yang lebih baik berbanding dengan kaedah-kaedah yang lain. Ketepatan keputusan yang menggalakkan adalah hasil daripada pemberat yang berpaduan dengan terma-terma sesuatu profil sewaktu proses pembelajaran.
Kajian ini membuka banyak ruang bagi kajian masa hadapan, termasuk pemeriksaan ke atas kebergantungan antara terma-terma tertentu dalam setiap profil, kualiti kaedah ini pada set data yang pelbagai, dan akhir sekali, penggunaan kaedah yang dicadangkan ini dalam bidang lain seperti sistem pencadangan.
ACKNOWLEDGEMENTS

In the name of Allah, the most Merciful, Gracious and most Compassionate, Creator of all knowledge for eternity. We beg for peace and blessing upon our Master the beloved Prophet Mohammed (Peace and Blessing be Upon Him) and his progeny, companions and followers.

I wish to extend my deepest appreciation and gratitude to the supervisory committee led by Assoc. Prof. Dr. Hj. Md. Nasir bin Sulaiman and committee member Pn. Hjh. Zaiton Muda for the various guidance, sharing of intellectual experiences and in giving me the vital to undertake the numerous of this study. Thanks to Pn. Aida Mustafa for her appreciated help.

I would like to thank Prof. Mohand Boughanem the Professor at Université Paul Sabatier of Toulouse, France for providing comments on the profile learning. I would also like to thank Lori R Price from Robotics Institute at Carnegie Mellon University, USA, for help with providing the AFC kit.

Also special thanks to my parents and my wife for making the best of my situation. May thanks are also extended to the Quraan teachers, my friends and the colleagues for sharing experiences throughout the years.
I certify that an Examination Committee has met on ..........2007 to conduct the final examination of Yahya Mohammed Al Murtadha on his Master thesis entitles “Improved Reinforcement-Based Profile Learning for Documents Filtering” in accordance with Universiti Pertanian Malaysia (Higher Degree) Act 1980 and Universiti Pertanian Malaysia (Higher Degree) Regulation 1981. The Committee recommends that the candidate be awarded the relevant degree. Members of the Examination Committee are as follows:

**Abdul Azim Abdul Ghani, PhD**
Associate Professor
Dean
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Chairman)

**Norwati Mustapha, PhD**
Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

**Masrah Azrifa Azmi Murad, PhD**
Lecturer
Faculty of Computer Science and Information Technology
Universiti Putra Malaysia
(Internal Examiner)

**Safaai Deris, PhD**
Professor
School of Graduate Studies
University Technology Malaysia
(External Examiner)

______________________________
HASANAH MOHD. GHAZALI, PhD
Professor / Deputy Dean
School of Graduate Studies
Universiti Putra Malaysia

Date:
This thesis submitted to the Senate of Universiti Putra Malaysia and has been accepted as fulfilment of the requirement for the degree of Master of Science. The members of the Supervisory Committee are as follows:

**Md Nasir Sulaiman, PhD**  
Associate Professor  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Chairman)

**Zaiton Muda, M.Sc.**  
Lecturer  
Faculty of Computer Science and Information Technology  
Universiti Putra Malaysia  
(Member)

---

**AINI IDERIS, PhD**  
Professor / Dean  
School of Graduate Studies  
Universiti Putra Malaysia  

Date: 13 August 2007
DECLARATION

I hereby declare that the thesis is based on my original work except for quotation and citation which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any degree at UPM or other institution.

YAHYA MOHAMMED AL MURTADHA

Date: 6 August 2007
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>DEDICATION</td>
<td>ii</td>
</tr>
<tr>
<td>ABSTRACT</td>
<td>iii</td>
</tr>
<tr>
<td>ABSTRAK</td>
<td>v</td>
</tr>
<tr>
<td>ACKNOWLEDGEMENTS</td>
<td>viii</td>
</tr>
<tr>
<td>APPROVAL</td>
<td>ix</td>
</tr>
<tr>
<td>DECLARATION</td>
<td>xi</td>
</tr>
<tr>
<td>LIST OF TABLES</td>
<td>xv</td>
</tr>
<tr>
<td>LIST OF FIGURES</td>
<td>xvi</td>
</tr>
<tr>
<td>LIST OF ABBREVIATIONS AND NOTATIONS</td>
<td>xviii</td>
</tr>
</tbody>
</table>

## CHAPTER

### 1 INTRODUCTION

1.1 Background                                                        1
1.2 Problem Statement                                                 4
1.3 Objective of the Research                                         6
1.4 Scope of the Research                                             6
1.5 Research Methodology                                              7
1.6 Contributions of the Research                                     8
1.7 Organization of the Thesis                                        9

### 2 LITERATURE REVIEW

2.1 Introduction                                                      12
2.2 Machine Learning                                                  13
   2.2.1 Reinforcement Learning                                        14
   2.2.2 Elements of Reinforcement Learning                            15
2.3 Information Filtering (IF)                                         16
   2.3.1 Major Filtering Methods                                      17
   2.3.2 Comparing IF and Information Retrieval (IR)                  19
2.4 Personalization                                                   20
   2.4.1 Content Tagging                                               21
   2.4.2 User Profiling                                                21
   2.4.3 Matchmaking                                                  23
   2.4.4 Personalization Research Directions                           24
2.5 Software Agents                                                    25
   2.5.1 The Agent Designing Concepts                                  26
   2.5.2 Intelligent Agents                                            27
   2.5.3 General-purpose Agent Development Suites                     29
   2.5.4 Multi Agent Systems (MAS) Models and Systems                 31
2.6 Previous Works on Personalized Information Filtering Agents       33
   2.6.1 MAS and Models Dealing with IF                                 34
   2.6.2 Drawbacks                                                     38
2.7 Related Works on Profile Learning Engines 39
  2.7.1 Rocchio Algorithm 40
  2.7.2 Query Expansion on Okapi/BSS Basic Search System 42
  2.7.3 Incremental Profile Leanring Using Reinforcement (Reinf.) 45
  2.7.4 Query Expansion 47
2.8 Summary 52

3 THE VECTOR SPACE MODEL VSM
  3.1 Introduction 53
  3.2 Vector Space Model (VSM) 53
  3.3 VSM Model Stages 57
    3.3.1 Document Indexing 57
    3.3.2 Term Weighting 58
    3.3.3 Similarity Coefficients 61
  3.4 VSM Illustration Example 64
  3.5 Discussion 67
  3.6 VSM Basic Model Improvements 68
  3.7 Documents Filtering using VSM 69
  3.8 Summary 70

4 RESEARCH METHODOLOGIES
  4.1 Introduction 72
  4.2 The Research K-Chart 72
  4.3 The Filtering Process Model 76
  4.4 RePLS Framework 77
  4.5 RePLS Inputs and Outputs 78
  4.6 Research Framework 79
  4.7 Text Retrieval Conference (TREC) 81
  4.8 TREC Track Used in This Thesis (Filtering Track) 83
    4.8.1 TREC DATA 85
    4.8.2 TREC Tasks 85
    4.8.3 Evaluation and Optimization 86
      4.8.3.1 Linear Utility 87
      4.8.3.2 Recall and Precision 88
  4.9 Summary 89

5 PROPOSED LEARNING PROFILE
  5.1 Introduction 91
  5.2 The Filtering Process 91
    5.2.1 The Indexing 92
    5.2.2 The Representation 95
    5.2.3 The Profile Manager 98
  5.3 Purity Terms Weighting 98
  5.4 Improved Profile Learning for Documents Filtering 101
  5.5 Profile Matching 105
6 AGENT-BASED PROFILE LEARNING FOR DOCUMENT FILTERING
6.1 Introduction 107
6.2 RETSINA: Carnegie Mellon University 108
6.3 The Agent Foundation Classes AFC 111
6.4 Agent Based RePLS Design 113
6.5 Scenario 117
6.6 Agent Based RePLS System Implementation 119
5.7 Summary 123

7 EXPERIMENTAL RESULTS AND DISCUSSIONS
7.1 Introduction 124
7.2 Experimental Remarks of the Proposed System 125
7.3 Discussion on the Purity Weighting Experiments 127
7.3.1 The Purity Occurrence in Irrelevant Documents Experiments 128
7.3.2 The Purity Occurrence in Relevant Documents Experiments 131
7.4 Discussion on the Experiments Results 138
7.5 The Evaluation 141
7.5.1 The Linear Utility 141
7.5.2 Recall and Precision 143
7.6 Discussion of the Proposed Agent Approach 144
7.7 Summary 148

8 CONCLUSIONS AND FUTURE WORKS
8.1 Overview 150
8.2 Concluding Remarks 153
8.3 Future Works and Extensions 155

REFERENCES 156
APPENDICES 163
BIODATA OF THE AUTHOR 207
# LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1</td>
<td>Agent-Based RePLS Scenario</td>
<td>118</td>
</tr>
<tr>
<td>6.2</td>
<td>Profile Manager Agent Descriptor</td>
<td>119</td>
</tr>
<tr>
<td>7.1</td>
<td>Training and Testing Document Set</td>
<td>126</td>
</tr>
<tr>
<td>7.2</td>
<td>Profile R101 Top Ten Terms Weights by Purity in Irrelevant Documents</td>
<td>128</td>
</tr>
<tr>
<td>7.3</td>
<td>Profile R101 Relevance Values for the Test Documents Set</td>
<td>129</td>
</tr>
<tr>
<td>7.4</td>
<td>Profile R101 Binary Relevance Values for Purity in Irrelevant Documents</td>
<td>130</td>
</tr>
<tr>
<td>7.5</td>
<td>Profile R101 T11SU using Purity in Irrelevant Documents</td>
<td>131</td>
</tr>
<tr>
<td>7.6</td>
<td>Profile R101 Top Ten Terms Weights using Purity in Relevant Documents</td>
<td>132</td>
</tr>
<tr>
<td>7.7</td>
<td>Profile R101 Relevance Values for the Test Documents Set</td>
<td>133</td>
</tr>
<tr>
<td>7.8</td>
<td>Profile R101 Binary Relevance Values for Purity in Relevant Documents</td>
<td>134</td>
</tr>
<tr>
<td>7.9</td>
<td>Profile R101 T11SU Evaluation</td>
<td>135</td>
</tr>
<tr>
<td>7.10</td>
<td>Ten Profiles Document Filtering using Four Methods</td>
<td>138</td>
</tr>
<tr>
<td>7.11</td>
<td>T11SU Evaluation of Ten Profiles</td>
<td>142</td>
</tr>
<tr>
<td>7.12</td>
<td>Recall and Precision for the Four Learning Methods</td>
<td>144</td>
</tr>
</tbody>
</table>
## LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>(Nwana's, 2003) Category of Software Agent</td>
<td>28</td>
</tr>
<tr>
<td>3.1</td>
<td>The VSM Representation Array</td>
<td>54</td>
</tr>
<tr>
<td>3.2</td>
<td>Terms and Document Vector Representation</td>
<td>55</td>
</tr>
<tr>
<td>3.3</td>
<td>The Representation of the Documents, Terms and the Queries in Three Dimensional Spaces</td>
<td>56</td>
</tr>
<tr>
<td>3.4</td>
<td>VSM Example (Distribution of Term “car” across Collection, Documents, Passages and Sentences)</td>
<td>61</td>
</tr>
<tr>
<td>3.5</td>
<td>Computing Similarity Scores</td>
<td>63</td>
</tr>
<tr>
<td>3.6</td>
<td>The Summary of Retrieval Results for the VSM Example</td>
<td>64</td>
</tr>
<tr>
<td>4.1</td>
<td>The Research K-Chart</td>
<td>75</td>
</tr>
<tr>
<td>4.2</td>
<td>The Filtering Model</td>
<td>76</td>
</tr>
<tr>
<td>4.3</td>
<td>RePLS Framework</td>
<td>78</td>
</tr>
<tr>
<td>4.4</td>
<td>RePLS Input and Output</td>
<td>79</td>
</tr>
<tr>
<td>4.5</td>
<td>Research Framework Flowchart</td>
<td>80</td>
</tr>
<tr>
<td>5.1</td>
<td>The Document Indexing as a “Bag of Words”</td>
<td>93</td>
</tr>
<tr>
<td>5.2</td>
<td>RePLS Document Representation in the Documents DB</td>
<td>97</td>
</tr>
<tr>
<td>5.3</td>
<td>Purity Learning Flowchart</td>
<td>104</td>
</tr>
<tr>
<td>6.1</td>
<td>RETSINA MAS</td>
<td>109</td>
</tr>
<tr>
<td>Section</td>
<td>Topic</td>
<td>Page</td>
</tr>
<tr>
<td>---------</td>
<td>-------</td>
<td>------</td>
</tr>
<tr>
<td>6.2</td>
<td>RETSINA Agent Architecture</td>
<td>111</td>
</tr>
<tr>
<td>6.3</td>
<td>Agent-Based RePLS</td>
<td>115</td>
</tr>
<tr>
<td>6.4</td>
<td>RePLS MAS</td>
<td>117</td>
</tr>
<tr>
<td>6.5</td>
<td>The Title and Description of Profile R101</td>
<td>119</td>
</tr>
<tr>
<td>6.6</td>
<td>The Agent Registration under ANS</td>
<td>120</td>
</tr>
<tr>
<td>6.7</td>
<td>The Agent Communication</td>
<td>120</td>
</tr>
<tr>
<td>6.8</td>
<td>RePLS Snapshot</td>
<td>122</td>
</tr>
<tr>
<td>6.9</td>
<td>An Example of RePLS Learning Output Database for R101</td>
<td>123</td>
</tr>
<tr>
<td>7.1</td>
<td>Purity Top Ten Terms for Profile R101</td>
<td>136</td>
</tr>
<tr>
<td>7.2</td>
<td>Profile R101 Mean’s Values of Selected Terms</td>
<td>137</td>
</tr>
<tr>
<td>7.3</td>
<td>Profile R101 Paired Samples Test for Purity-Reinf and Purity-Rocchio</td>
<td>137</td>
</tr>
<tr>
<td>7.4</td>
<td>Purity Proportion of Number of Selected Relevant Documents to Number of Relevant Documents</td>
<td>140</td>
</tr>
<tr>
<td>7.5</td>
<td>Proportion of Number of Selected Relevant Documents to Number of Relevant Documents Using Four Methods</td>
<td>141</td>
</tr>
<tr>
<td>7.6</td>
<td>The Relationship between T11SU and the Number of Relevant Documents</td>
<td>143</td>
</tr>
<tr>
<td>7.7</td>
<td>RePLS Function Diagram</td>
<td>146</td>
</tr>
<tr>
<td>7.8</td>
<td>PMA Term’s Explanation</td>
<td>147</td>
</tr>
</tbody>
</table>
# LIST OF ABBREVIATIONS

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>AFC</td>
<td>Agent Foundation Classes</td>
</tr>
<tr>
<td>ANS</td>
<td>Agent Name Server</td>
</tr>
<tr>
<td>IDF</td>
<td>Inverse Document Frequency</td>
</tr>
<tr>
<td>IF</td>
<td>Information Filtering</td>
</tr>
<tr>
<td>IR</td>
<td>Information Retrieval</td>
</tr>
<tr>
<td>MAS</td>
<td>Multi Agent System</td>
</tr>
<tr>
<td>ML</td>
<td>Machine Learning</td>
</tr>
<tr>
<td>NIST</td>
<td>National Institute of Standards and Technology</td>
</tr>
<tr>
<td>Okapi/BSS</td>
<td>Okapi Basic Search System</td>
</tr>
<tr>
<td>ODBC</td>
<td>Open Database Connectivity</td>
</tr>
<tr>
<td>PMA</td>
<td>Profile Manager Agent</td>
</tr>
<tr>
<td>RCV</td>
<td>Reuters Corpus Volume</td>
</tr>
<tr>
<td>RePLS</td>
<td>Reinforcement Profile Learning System</td>
</tr>
<tr>
<td>Reinf.</td>
<td>Reinforcement Incremental Profile Learning</td>
</tr>
<tr>
<td>SA</td>
<td>Software Agents</td>
</tr>
<tr>
<td>TREC</td>
<td>Text Retrieval Conference</td>
</tr>
<tr>
<td>VSM</td>
<td>Vector Space Model</td>
</tr>
<tr>
<td>XML</td>
<td>Extended Markup Language</td>
</tr>
</tbody>
</table>
CHAPTER 1

INTRODUCTION

Background

There are numerous text documents available in electronic form. More and more are becoming available every day. Such documents represent a massive amount of information. In addition, the information sources set a dynamic and unorganized environment where the information appear and disappear at any time. Gathering information from such environment is similar to drinking water from a fire hose metaphorically. Hence, there are many occasions when users are not able to get the information they require. Seeking values in this huge collection requires organization, and much of the work of organizing documents can be automated through text classification (Rennie, 2001).

To alleviate this problem, the solution of Information Filtering (IF) and user profiling is introduced. Filters are tools to help people find the most valuable information, so that the limited time spent on locating the information can be maximized on finding the most interesting and valuable documents. Filters are also used to organize and structure information. IF monitors the incoming documents and filter only those matches the user’s information need, which are called profile. A typical case is a newsfeed, where the producer is a news agency.
and the consumer is a newspaper. In this case, the filtering system should block the delivery of the documents that readers are not interested in (e.g., all news not concerning sports, in the case of a sports newspaper). Filtering is done by applying filtering rules to attributes of the documents to be filtered. The attributes of these documents, are mainly words in the titles, abstracts, or the whole document. Filtering can be seen as a case of single-label Text Categorization, which is the classification of incoming documents into two disjoint categories: the relevant and the irrelevant.

The uncertainties in the filtering environments especially the dynamic nature of the user’s interests and the documents stream have made it extremely difficult to gather and maintain accurate information necessary for filtering (Mostafa et al., 1997). Viewed from the perspective of the filtering system, rapid changes introduced in the environment are sources of uncertainty. Managing such uncertainties require a high level of adaptivity on the system’s part. This adaptivity can be achieved by applying various machine learning techniques. The overall problem of the IF may then be interpreted as learning the mapping from a space of documents to the space of real-valued user relevance factors.

Profile learning is the heart of a filtering process. It aims to collect the user preferences and modify the filtering behavior according to the preferences. To identify whether a document is relevant to the user or not, a score that measures the similarity between the document and the profile is computed. When the score
is higher than the similarity threshold then the document is selected, otherwise the document is rejected. Profile learning has been studied by two communities, Information Filtering and Machine Learning (Tebri et al., 2005). In machine Learning (ML) paradigm, a general inductive process will automatically build an automatic document classifier by learning from a set of pre-classified documents. The advantages of this approach are; (1) the accuracy is comparable to the accuracy achieved by human experts, and (2) a considerable saving in terms of expert labor power, since no intervention from either the knowledge engineers or the domain experts is needed for construction of the classifier or for porting to a different set of categories (Sebastian, 2002).

Most of the IF researches is based on the Rocchio's algorithm (Rocchio, 1971). It is based on the Vector Space Model (VSM) methodology that allows for partial matching by assigning non-binary (as opposed to the Boolean method) weights to index terms in both profiles and documents. These weights indicate the importance of the terms in describing the semantic of the document and profiles. Terms weights are used to calculate the degree of similarity between each document stored in the system and the user profiles. This method however, does not cover text mining, which has another focus compared to text retrieval (Boertjes et al., 2001). In general, text mining aims at finding implicit correlations in texts by trying to discover previously unknown information. Text retrieval on the other hand, focuses on finding information that is already present explicitly in texts.
IF systems must acquire and maintain accurate knowledge regarding the documents as well as the users. The dynamic nature of the user interests and the document streams makes the maintenance of such knowledge quite complex. Acquiring correct user interest profiles is difficult; users may not be sure of their interests or even do not wish to invest an effort in creating such a profile. Acquiring information regarding documents is equally difficult, due to the size of the document stream and the computational cost associated with parsing huge texts. At any time, new topics may be introduced in the document stream, or user interests related to topics may change. Furthermore, a sufficiently represented document may not be available to facilitate a prior analysis or training.

**The Problem Statement**

As the amount of accessible information is overwhelming, an intelligent and personalized filtration of available information is a big challenge (Albayrak et al., 2005). The accumulative number of information containers, the users and their increasing demands for the information are the main cause of the retrieval of huge and irrelevant information. IF monitors a stream of incoming document to find those match the user’s information need, known as the profile. Profile learning is fundamental in the filtering process; the goal is to collect the user preference on the judged documents and modify the filtering behavior accordingly.
The capability of modeling and learning the user preferences is at the heart of a personalized information filtering system. The main problem with all the personalized filtering is how to measure and select the most suitable terms (attributes) that will help to discriminate between the filtered classes, and learn the user’s interests to build their profiles.

Most of the researches in IF use an incremental version of Rocchio's algorithm (Salton and Buckley, 1990; Singhal et al., 1997) to propose different profile learning methods. Microsoft Research Laboratory in Cambridge has developed an evaluation environment called Keenbow for a wide range of IR experiments. One component in Keenbow is Okapi/BSS (Robertson and Walker[a], 2000) which uses terms weighting in addition to the Query Expansion. Reinf, the incremental profile learning based on reinforcement method (Tebri et al., 2005) is an IF profile learning method based on the terms weighting. These methods weigh the terms according to the frequency of term occurrence in the documents and the profiles, without considering the pure occurrence in either the relevant or irrelevant documents.

With the growing need for a sufficient learning of user interests, there also exists a growing urgency to achieve this without an additional effort from the user by adapting the software agent approach. Software agents are software entities that
array out some set of operations on behalf of a user or another program with some degree of independence or autonomy (IBM Agent\(^1\)).

**Objectives of the Research**

The main goal of this research is to improve the document filtering based on user interests. To fulfill this, the following objectives must be achieved:

- Improving the filtering engine at the terms weighting stage of the reinforcement-based profile learning, in order to select the best terms to aid for building the user profiles based on the user preference.
- Implementing an agent-based document filtering system that works as a user assistant to help building the user profiles based on his preferences.

**Scope of the Research**

This research focuses on improving terms weighting for learning the user preferences and building the user profile. The learning method is an incremental profile learning based on the reinforcement method. A user profile learning agent that is able to help learning the user preferences without any user intervention is also implemented. The data used for testing is TREC 2002 filtering track, which is a Reuter’s news stories in XML format. The output of the learning stage is a list of selected terms for each user profile stored in database tables. The list is then