Cognitive Intelligence and Robotics

Series editors

Amit Konar, Department of Electronics and Tele-communication Engineering, Jadavpur University, Kolkata, India
Witold Pedrycz, Department of Electrical and Computer Engineering, University of Alberta, Edmonton, AB, Canada
Cognitive Intelligence refers to the natural intelligence of humans/animals involving the brain to serve the necessary biological functioning to perform an intelligent activity. Although tracing a hard boundary to distinguish intelligent activities from others remains controversial, most of the common behaviors/activities of living organisms that cannot be fully synthesized by artificial means are regarded as intelligent. Thus the act of natural sensing and perception, understanding of the environment and voluntary control of muscles, blood-flow rate, respiration rate, heartbeat, and sweating rate, which can be performed by lower level mammals, indeed, are intelligent. Besides the above, advanced mammals can perform more sophisticated cognitive tasks, including logical reasoning, learning and recognition and complex planning/coordination, none of which could be realized artificially to the level of a baby, and thus are regarded as cognitively intelligent.

The series aims at covering two important aspects of the brain science. First, it would attempt to uncover the mystery behind the biological basis of cognition with special emphasis on the decoding of stimulated brain signals/images. The coverage in this area includes neural basis of sensory perception, motor control, sensory-motor coordination and also understanding the biological basis of higher-level cognition, such as memory and learning, reasoning and complex planning. The second objective of the series is to publish brain-inspired models of learning, perception, memory and coordination for realization on robots to enable them to mimic the cognitive activities performed by the living creatures. These brain-inspired models of machine intelligence would supplement the behavioral counterparts, studied in traditional AI.

The series includes textbooks, monographs, contributed volumes and even selected conference proceedings.

More information about this series at http://www.springer.com/series/15488
Debajyoti Mukhopadhyay
Editor

Web Searching and Mining

Springer
Preface

Searching the Web has become a natural process of our day-to-day life. Even a non-technical person with very little technical knowledge searches the Web by virtue of owning a cell phone handset with Internet connectivity. Mining the data follows as a part of the necessity to pick the best-suited data. However, this book entitled *Web Searching and Mining* is not meant for a Web-searching enthusiast. Rather, it is more of a research-based book, exploring the new possibilities of introducing the power of cellular automata theory in the field of search engines, which helps reduce significantly the storage of data space. It also brings in hands-on experience of utilizing ontology to store and search data of specific domains.

While guiding my research scholars in the related field, Anirban Kundu and Sukanta Sinha went on to earn their Ph.D. degrees in these related fields and their work created the base of this book.

During 1982–1994, while in the USA, I had the chance to work in the field of distributed computing at Bell Communications Research in New Jersey. After returning to India, I felt the need of setting up a research laboratory combining the power of distributed computing with the emerging field of Web technology. As a result of that effort, Web Intelligence and Distributed Computing Research Lab (WIDiCoReL) was set up in Kolkata in 2002 and several research scholars had carried out their B.E., M.E. and Ph.D. works under the aegis of WIDiCoReL. In this laboratory, we have explored the power of cellular automata and, perhaps for the first time in the research history in the globe, introduced cellular automata in the field of Web searching, Web mining, Web-page prediction, etc. A large number of publications came out of this research initiative and are available on the Internet.

I take this opportunity to extend my sincere thanks to my former research scholars, Dr. Anirban Kundu and Dr. Sukanta Sinha, for making the resources available while writing this book. Sincere thanks to the team at Springer for making the book a reality.

Kolkata, India

Dr. Debajyoti Mukhopadhyay
Web Intelligence and Distributed Computing Research Lab
Contents

Introduction .............................................................................................................. 1
Debajyoti Mukhopadhyay and Sukanta Sinha

Preliminaries on Cellular Automata ................................................................. 29
Debajyoti Mukhopadhyay and Anirban Kundu

Design of SMACA ............................................................................................... 37
Debajyoti Mukhopadhyay and Anirban Kundu

SMACA Usage in Indexing Storage of a Search Engine .............................. 51
Debajyoti Mukhopadhyay and Anirban Kundu

Cellular Automata in Web-Page Ranking ....................................................... 65
Debajyoti Mukhopadhyay and Anirban Kundu

Web-Page Indexing Based on the Prioritize Ontology Terms ..................... 75
Debajyoti Mukhopadhyay and Sukanta Sinha

Domain-Specific Crawler Design ................................................................. 85
Debajyoti Mukhopadhyay and Sukanta Sinha

Structural Change of Domain-Specific Web-Page Repository for Efficient Searching ......................................................... 113
Debajyoti Mukhopadhyay and Sukanta Sinha

Domain-Specific Web-Page Prediction ......................................................... 145
Debajyoti Mukhopadhyay and Sukanta Sinha

Domain-Specific Lucky Searching ............................................................... 155
Debajyoti Mukhopadhyay and Sukanta Sinha
About the Editor

Dr. Debajyoti Mukhopadhyay is currently Director and Dean (R&D) at NHITM affiliated to Mumbai University, India. He previously worked in the IT industry for 19 years, including at the well-known Bell Communications Research, USA, and in academia for 16 years, including as the Dean (R&D) of Maharashtra Institute of Technology, Pune, India. He has published over 190 research papers and three patents. He previously worked in the corporate sector, holding top-level positions, such as the president, CEO, director, and general manager, and oversaw a large number of professionals managing multiple offshore projects from India. He has been elected as Distinguished Speaker of the Computer Society of India. He had held visiting positions at Chonbuk National University, South Korea; George Mason University, USA; and Thapar University, India. He holds a Ph.D. in engineering from Jadavpur University, India; an M.S. in computer science from Stevens Institute of Technology, USA; a postgraduate diploma in computer science from the Queen’s University of Belfast, UK; and a B.E. in electronics and telecommunications engineering from Bengal Engineering College under the University of Calcutta, India. He is FIE, FIETE, SMIEEE, and SMACM, USA; Chartered Engineer, MIMA (India), and Elected Member of Eta Kappa Nu (the EE Honor Society of the USA).
List of Figures

Introduction
Fig. 1 Basics of Web search engine components ................... 4
Fig. 2 Standard crawling ........................................ 7
Fig. 3 Domain-specific crawling .................................. 8
Fig. 4 A part of computer science ontology ...................... 10
Fig. 5 Domain-specific Web-page repository ..................... 11
Fig. 6 a Regular lucky searching, b Domain-specific lucky searching ........................................ 14
Fig. 7 a A snapshot of google search result, b A snapshot of yahoo search result, c A snapshot of MSN (bing beta) search result .......................... 16
Fig. 8 Image search for a popular image (Sachin Tendulkar) ........ 19
Fig. 9 Image search for a non-popular image (xyz) ................. 20

Preliminaries on Cellular Automata
Fig. 1 Local interactions between cellular automata cells ........... 30
Fig. 2 State transition behavior of cellular automata rule vector <6 240 60 65> ............................................... 31
Fig. 3 Example of grid structures of CA cells ........................ 32

Design of SMACA
Fig. 1 Structure of a SMACA with RV <4 102 53 85> ................ 39
Fig. 2 RVG and related RMT of a 4-cell CA with RV <2, 44, 4, 21> ................................................................. 44
Fig. 3 RVG and related RMT of a 5-cell CA with RV <13, 110, 237, 201, 81> .......................................................... 45
Fig. 4 RVG and related RMT of a 4-cell CA with RV <7, 131, 175, 65> ................................................................. 49

SMACA Usage in Indexing Storage of a Search Engine
Fig. 1 Schematic diagram of a typical search engine ................. 53
Fig. 2 Pictorial representation of our approach ..................... 54
<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3</td>
<td>Space required for forward indexing</td>
<td>61</td>
</tr>
<tr>
<td>4</td>
<td>Space required for inverted indexing</td>
<td>62</td>
</tr>
</tbody>
</table>

**Cellular Automata in Web-Page Ranking**

<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Pictorial view of in-link and out-link connections of web-pages</td>
<td>67</td>
</tr>
<tr>
<td>2</td>
<td>Structure of GF ( (2^k) ) CA</td>
<td>68</td>
</tr>
<tr>
<td>3</td>
<td>Structured model of forward indexing and inverted indexing.</td>
<td>72</td>
</tr>
</tbody>
</table>

**Web-Page Indexing Based on the Prioritize Ontology Terms**

<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Example of extracting dominating and sub-dominating Ontology terms</td>
<td>77</td>
</tr>
<tr>
<td>2</td>
<td>Web-page structures after applying our indexing mechanism</td>
<td>79</td>
</tr>
<tr>
<td>3</td>
<td>A part of user interface</td>
<td>80</td>
</tr>
</tbody>
</table>

**Domain-Specific Crawler Design**

<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Relevance calculation of a Web-page</td>
<td>88</td>
</tr>
<tr>
<td>2</td>
<td>Checking domain of a Web-page</td>
<td>89</td>
</tr>
<tr>
<td>3</td>
<td>Challenge in our approach</td>
<td>89</td>
</tr>
<tr>
<td>4</td>
<td>Relevant page tree from original crawling</td>
<td>90</td>
</tr>
<tr>
<td>5</td>
<td>RPaT searching</td>
<td>91</td>
</tr>
<tr>
<td>6</td>
<td>Multiple domains specific crawling</td>
<td>92</td>
</tr>
<tr>
<td>7</td>
<td>Multiple domains relevance calculation of a Web-page</td>
<td>93</td>
</tr>
<tr>
<td>8</td>
<td>Checking multiple domains specific Web-pages</td>
<td>95</td>
</tr>
<tr>
<td>9</td>
<td>Arbitrary example of a RPaG</td>
<td>96</td>
</tr>
<tr>
<td>10</td>
<td>RPaG reading mechanism</td>
<td>97</td>
</tr>
<tr>
<td>11</td>
<td><strong>a</strong> Web-page content classifier, <strong>b</strong> Web-page URL classifier</td>
<td>98</td>
</tr>
<tr>
<td>12</td>
<td>A part of user interface</td>
<td>99</td>
</tr>
<tr>
<td>13</td>
<td>Proposed architecture of domain-specific Web search engine resource collector</td>
<td>99</td>
</tr>
<tr>
<td>14</td>
<td>Harvest rate for unfocused crawling</td>
<td>103</td>
</tr>
<tr>
<td>15</td>
<td>Harvest rate of focused crawler</td>
<td>103</td>
</tr>
<tr>
<td>16</td>
<td>Harvest rate for focused crawling with tolerance limit 10 and relevance limit 5</td>
<td>104</td>
</tr>
<tr>
<td>17</td>
<td>Domainwise Web-page distribution</td>
<td>107</td>
</tr>
<tr>
<td>18</td>
<td>Time taken in single domain crawling and multiple domains crawling</td>
<td>107</td>
</tr>
<tr>
<td>19</td>
<td>Line by line complexity analysis</td>
<td>109</td>
</tr>
<tr>
<td>20</td>
<td>Harvest rate for focused crawling</td>
<td>109</td>
</tr>
</tbody>
</table>

**Structural Change of Domain-Specific Web-Page Repository for Efficient Searching**

<table>
<thead>
<tr>
<th>Fig.</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>High-efficient relevant page tree (HERT).</td>
<td>114</td>
</tr>
<tr>
<td>2</td>
<td>HERT searching</td>
<td>115</td>
</tr>
<tr>
<td>3</td>
<td>Dummy pages for HERT initialization</td>
<td>116</td>
</tr>
<tr>
<td>4</td>
<td>HERT after insertion of Web-page ( a )</td>
<td>118</td>
</tr>
</tbody>
</table>
List of Figures

Fig. 5   HERT after insertion of Web-page b and c .......................... 118
Fig. 6   HERT after insertion of Web-page d ................................. 119
Fig. 7   HERT after insertion of Web-page e ................................. 119
Fig. 8   HERT after insertion of Web-page f ................................. 119
Fig. 9   HERT after insertion of Web-page g ................................. 120
Fig. 10  HERT after insertion of Web-page h and Web-page i ............. 120
Fig. 11  HERT after insertion of Web-page j and Web-page k .............. 120
Fig. 12  HERT after insertion of Web-page l ................................ 121
Fig. 13  HERT after insertion of Web-page m ................................ 121
Fig. 14  HERT after insertion of Web-page n ................................ 122
Fig. 15  HERT after insertion of all RPaT Web-pages ........................ 122
Fig. 16  Index-based acyclic graph (IBAG) .................................... 123
Fig. 17  Sample IBAG Web-page searching .................................... 125
Fig. 18  Dummy IBAG ............................................................ 128
Fig. 19  IBAG after Web-page 'a' insertion .................................... 128
Fig. 20  Final IBAG from given RPaG .......................................... 129
Fig. 21  A part of user interface ................................................ 130
Fig. 22  Selected pages in IBAG ................................................ 131
Fig. 23  IBAG (Ideal Case) ....................................................... 132
Fig. 24  IBAG (while \((n/m) \approx n\)) ......................................... 132
Fig. 25  M-IBAG Model .......................................................... 133
Fig. 26  HERT page storage distribution ....................................... 137
Fig. 27  Time taken in HERT searching and RPaT searching .......... 138

Domain-Specific Web-Page Prediction
Fig. 1   Comparison between average number of web-pages Retrieved from before and after bit masking in IBAG model ............ 150
Fig. 2   Comparison between average time taken for searching web-pages from before and after bit masking in IBAG model. . 153

Domain-Specific Lucky Searching
Fig. 1   DSLSDB construction .................................................. 158
Fig. 2   Lucky URL retrieve .................................................... 159
Fig. 3   A part of user interface ................................................ 160
Fig. 4   Web-page content of a domain-specific page .................. 162
Fig. 5   Error message for invalid search string ........................ 164
List of Tables

**Introduction**
Table 1  Basic information about each product .......................... 17

**Preliminaries on Cellular Automata**
Table 1  Truth table of sample rules of a CA cell showing the next state logic for the minterms of a 3 variable boolean function—The 8 minterms having decimal values 0, 1, 2, 3, 4, 5, 6, 7 are referred to as rule minterms (RMTS) ..................... 30
Table 2  Next state logic of a few rules .......................... 31
Table 3  Linear/additive CA rules employing next state function with XOR/XNOR logic. ............................... 32

**Design of SMACA**
Table 1  SMACA class table. ................................... 40
Table 2  Relationship of \( R_i \) and \( R_{i+1} \) ..................... 41
Table 3  First rule table \( (R_0) \) ................................... 42
Table 4  Last rule table \( (R_{n-1}) \) ............................. 42
Table 5  Generic node values of RVG for even distribution of RMTs on edges ................................... 46
Table 6  Probable next level RMTs ............................... 46
Table 7  Generic node values of RVG for uneven distribution of RMTs on edges ................................... 46

**SMACA Usage in Indexing Storage of a Search Engine**
Table 1  Experimental results on time required for searching ......... 63

**Web-Page Indexing Based on the Prioritize Ontology Terms**
Table 1  Performance report of our system ......................... 82
Table 2  Accuracy of our system .................................. 82
Domain-Specific Crawler Design
Table 1 RPaT page repository ........................................ 91
Table 2 Sample seed URLs for single domain-specific web
search crawler .................................................. 102
Table 3 Sample weight table for computer science domain .......... 102
Table 4 Sample weight table for computer science domain .......... 104
Table 5 Sample seed URLs for multiple domains specific Web search
crawler .......................................................... 105
Table 6 Sample synterms for cricket ontology ....................... 105
Table 7 Sample synterms for football ontology ..................... 105
Table 8 Sample synterms for hockey ontology ..................... 106
Table 9 Sample weight table structure for cricket ontology ......... 106
Table 10 Sample weight table structure for football ontology ...... 106
Table 11 Sample weight table structure for hockey ontology ...... 106
Table 12 Performance report of multilevel domain-specific crawler .. 110

Structural Change of Domain-Specific Web-Page Repository for
Efficient Searching
Table 1 RANGE_INDEX table ..................................... 115
Table 2 Seed URLs .................................................. 135
Table 3 Weight table ............................................... 136
Table 4 Syntable .................................................... 136
Table 5 Comparative study of time complexity ...................... 143

Domain-Specific Web-Page Prediction
Table 1 Accuracy measure statistics ................................ 151

Domain-Specific Lucky Searching
Table 1 Seed URLs .................................................. 161
Table 2 Weight table ............................................... 161
Table 3 Syntable .................................................... 162
Table 4 Comparative study of Lucky search engines ............... 164