

**Intelligent Web Pre-Fetching Method**

by

**O Kit Hong**

**Master of Science in Software Engineering**

**1999**



**Faculty of Science and Technology  
University of Macau**

# Table of Contents

<b>List of Figures</b>	<b>ix</b>
<b>List of Tables</b>	<b>xi</b>
<b>1 Introduction</b>	<b>1</b>
<b>2 Current Research on Pre-fetching</b>	<b>3</b>
2.1 Server-Initiated Pre-fetching . . . . .	3
2.2 Client-Initiated Pre-fetching . . . . .	6
2.2.1 Deterministic Client-Initiated Pre-fetching . . . . .	6
2.2.2 Statistical Client-Initiated Pre-fetching . . . . .	7
2.3 Proxy-Initiated Pre-fetching . . . . .	9
2.3.1 Lookahead Proxy-Initiated Pre-fetching . . . . .	9
2.4 Hierarchical Pre-fetching . . . . .	10
2.4.1 Server to Server Hierarchical Pre-fetching . . . . .	11
2.4.2 Client to Server Hierarchical Pre-fetching . . . . .	12
2.5 Summary . . . . .	14
<b>3 A New Pre-fetching Model</b>	<b>17</b>
3.1 System Overview . . . . .	17
3.1.1 Proxy-Initiated Pre-fetching . . . . .	18
3.2 System Architecture . . . . .	18
3.2.1 User Access Log Profile . . . . .	19
3.2.2 Periodical User's Interest Profile . . . . .	20
3.2.3 Global Information Profile . . . . .	21
3.2.4 Current Interest Register . . . . .	21
3.3 Prediction Algorithm . . . . .	22
3.3.1 Lookahead Pre-fetching . . . . .	22
3.3.2 Keyword-Based Prediction . . . . .	22
3.3.3 Algorithm . . . . .	23
3.4 Performance and Experimental Considerations . . . . .	23

3.4.1	Scheduling Requests . . . . .	23
3.4.2	Parameterized Thresholds and Adjustable Parameters . . . . .	24
3.4.3	Pre-fetching HTML File Only . . . . .	24
3.5	Special Technologies Required for This Model . . . . .	24
3.5.1	Keyword Extraction . . . . .	25
3.5.2	User Interest Analysis . . . . .	26
3.5.3	User Identification . . . . .	27
3.5.4	Knowledge of Unaccessed Web Pages . . . . .	28
<b>4</b>	<b>Keyword Extraction</b>	<b>29</b>
4.1	Limitation on Traditional Methods . . . . .	29
4.2	The Possibility of Using HTML Structure . . . . .	31
4.3	Integrated Method Used in the New Proposed Model . . . . .	33
4.3.1	Keyword Extraction . . . . .	33
4.3.2	Keyword Weighting . . . . .	35
4.3.3	Experiment . . . . .	35
4.3.4	Vector Representation and Similarity Computation . . . . .	39
4.4	Configurable Parameters . . . . .	42
4.5	Different Conventions and Linguistic . . . . .	42
<b>5</b>	<b>User Interest Analysis</b>	<b>43</b>
5.1	User Access Log Profile Pattern . . . . .	43
5.2	Periodical User's Interest Profile Pattern . . . . .	44
5.3	Global Information File Pattern . . . . .	45
5.4	Current Interest Register . . . . .	46
5.5	Determining the User's Current Interest . . . . .	46
5.5.1	Deflating Factors . . . . .	46
5.5.2	Start-Up Calculation . . . . .	48
5.5.3	Calculation on the Fly . . . . .	48
5.5.4	Periodical User's Interest Profile — Contribution and Feedback . . . . .	48
5.5.5	Configurable Parameters . . . . .	49
5.5.6	Furture Experimentations . . . . .	51
<b>6</b>	<b>User Identification</b>	<b>53</b>
6.1	Identifying a User by IP Address . . . . .	53
6.2	Identifying a Web Session . . . . .	54
<b>7</b>	<b>Conclusion</b>	<b>57</b>
7.1	Summary of Achievements . . . . .	57
7.2	Future Work . . . . .	58

<b>A</b>	<b>Program Source</b>	<b>63</b>
A.1	HTML Parser . . . . .	63
A.2	Lovins Stemmer . . . . .	66
A.3	Similarity Calculator . . . . .	84
<b>B</b>	<b>The Content of Web Pages</b>	<b>87</b>
B.1	URL: <a href="http://www.imap.org/whatisIMAP.html">http://www.imap.org/whatisIMAP.html</a> . . . . .	88
B.2	URL: <a href="http://squid.nlanr.net/Squid/">http://squid.nlanr.net/Squid/</a> . . . . .	89
B.3	URL: <a href="http://www.hbuk.co.uk/ap/journals/hu/">http://www.hbuk.co.uk/ap/journals/hu/</a> . . . . .	90
B.4	URL: <a href="http://www.real.com/publisher/index.html">http://www.real.com/publisher/index.html</a> . . . . .	92
B.5	URL: <a href="http://www.real.com/publisher/presenter.html">http://www.real.com/publisher/presenter.html</a> . . . . .	94
B.6	URL: <a href="http://healthyideas.com/weight/gym/">http://healthyideas.com/weight/gym/</a> .. . . .	96
B.7	URL: <a href="http://www.cnn.com/WORLD/africa/9808/">http://www.cnn.com/WORLD/africa/9808/</a> .. . . .	97
B.8	URL: <a href="http://www.info.gov.hk/info/iprights.htm">http://www.info.gov.hk/info/iprights.htm</a> . . . . .	101
B.9	URL: <a href="http://www.fst.umac.mo/staff/fstrpba.html">http://www.fst.umac.mo/staff/fstrpba.html</a> . . . . .	104
B.10	URL: <a href="http://station.nasa.gov/core.html">http://station.nasa.gov/core.html</a> . . . . .	106
B.11	URL: <a href="http://www.traderonline.com/auto/">http://www.traderonline.com/auto/</a> . . . . .	107
B.12	URL: <a href="http://www.well.com/user/kr2/">http://www.well.com/user/kr2/</a> . . . . .	108
B.13	URL: <a href="http://www.dealer.com/">http://www.dealer.com/</a> . . . . .	116
B.14	URL: <a href="http://www.edmunds.com/edweb/used/">http://www.edmunds.com/edweb/used/</a> .. . . .	117
B.15	URL: <a href="http://www.realnames.com/Resolver.dll">http://www.realnames.com/Resolver.dll</a> .. . . .	118
B.16	URL: <a href="http://www.netauto.com/">http://www.netauto.com/</a> . . . . .	119
B.17	URL: <a href="http://www.NorthsideAuto.com/">http://www.NorthsideAuto.com/</a> . . . . .	121