Matthias Eichstädt
Internet Webcasting

Matthias Eichstädt

Internet Webcasting

Generating and Matching Profiles

Springer Fachmedien Wiesbaden GmbH

Eichstädt, Matthias:

Internet webcasting : generating and matching profiles / Matthias Eichstädt.

(DUV : Informatik)

Zugl.: Hagen, Fernuniv., Diss., 1999

ISBN 978-3-8244-2125-1

ISBN 978-3-663-10402-5 (eBook)

DOI 10.1007/978-3-663-10402-5

Alle Rechte vorbehalten

© Springer Fachmedien Wiesbaden 1999

Ursprünglich erschienen bei Deutscher Universitäts-Verlag GmbH, Wiesbaden, 1999

Lektorat: Claudia Splittgerber / Monika Mülhausen

Das Werk einschließlich aller seiner Teile ist urheberrechtlich geschützt. Jede Verwertung außerhalb der engen Grenzen des Urheberrechtsgesetzes ist ohne Zustimmung des Verlages unzulässig und strafbar. Das gilt insbesondere für Vervielfältigungen, Übersetzungen, Mikroverfilmungen und die Einspeicherung und Verarbeitung in elektronischen Systemen.

<http://www.duv.de>

Höchste inhaltliche und technische Qualität unserer Produkte ist unser Ziel. Bei der Produktion und Verbreitung unserer Bücher wollen wir die Umwelt schonen. Dieses Buch ist deshalb auf säurefreiem und chlorfrei gebleichtem Papier gedruckt. Die Einschweißfolie besteht aus Polyäthylen und damit aus organischen Grundstoffen, die weder bei der Herstellung noch bei der Verbrennung Schadstoffe freisetzen.

Die Wiedergabe von Gebrauchsnamen, Handelsnamen, Warenbezeichnungen usw. in diesem Werk berechtigt auch ohne besondere Kennzeichnung nicht zu der Annahme, daß solche Namen im Sinne der Warenzeichen- und Markenschutz-Gesetzgebung als frei zu betrachten wären und daher von jedermann benutzt werden dürften.

This thesis is dedicated to my family.

Acknowledgements

I would like to thank Professor Dr. Gunter Schlageter and Dr. Qi Lu for advising this dissertation. They have been an endless source of knowledge, encouragement, guidance and energy. I could not have had better mentors. They always made time to see me, no matter how busy their schedules were. Many of my researches have been following their foot steps.

I would like to thank Professor Dr. Shang-Hua Teng for serving as a member on my committee and for many fruitful discussions that have benefited me a lot. He taught me the importance of mathematical models in algorithm design and more importantly he has been a true and steady friend.

I could not have completed this thesis without the care and support of many wonderful people and it is an honor for me to acknowledge them here. I warmly thank my colleagues in the GCS group: Norm Pass, Daniel Ford, Joshua Dobies, Joseph Gebis, Reiner Kraft, Peter Lazarus, Toby Lehman, Udi Manber, Wayne Niblack, Ray Strimaitis, Neel Sundaresan, John Thomas, and Peter Yim. They are all very talented individuals and it was a true pleasure to work with them.

This dissertation was written while working at the IBM Almaden Research Center in San Jose, California. It was presented to the faculty of Computer Science at the FernUniversität Hagen in candidacy for the degree of Doctor of Philosophy. This work was partially sponsored by funds from IBM's Research Division. Even though specific IBM products are mentioned in this document, no conclusions should be drawn about future IBM product plans based on this publication's contents. The opinions expressed here are our own.

After I had completed the research presented in this thesis I joined Yahoo! Inc. in Santa Clara, California where I can be reached via electronic mail (matthias@yahoo-inc.com).

Matthias Eichstaedt

Contents

1	Introduction	1
1.1	Internet Webcasting	1
1.2	Personalization	2
1.3	Main Contributions	3
1.4	Organization of this Document	4
2	The IBM Grand Central Station Project	5
2.1	Introduction	5
2.2	Overview	6
2.3	New Frontiers in Information Dissemination	7
2.4	System Architecture	8
2.4.1	Data Collection	9
2.4.2	Data Distribution	10
3	Profile Language	13
3.1	Design Objectives and Constraints	13
3.2	Related Research	15
3.3	The Query Algebra	16
3.3.1	Profile Syntax	16
3.3.2	Predicate Syntax	17
4	Internal Profile Representation	23
4.1	Design Rationale	23
4.2	Data Structure	24
4.3	Profile Index Maintenance	25
4.4	Eliminating Redundant Nodes	27
4.5	Loading Large Profile Collections	28

5	Sequential Profile Matching	33
5.1	The Profile Matching Problem	33
5.2	Design Objectives	34
5.3	Profile Evaluation	34
5.4	Cost/Credit Based Leaf Node Ranking	37
5.4.1	Design Rationale	37
5.4.2	Ranking Algorithm	37
5.4.3	Initialization	39
5.5	Optimizations and Extensions	39
5.5.1	Ranking Leaf Nodes Beyond History	39
5.5.2	Group Evaluation of Leaf Nodes	41
5.5.3	Evaluating Indexed Catalog	41
6	Parallel Profile Matching	45
6.1	Design Objectives	45
6.2	Parallel Profile Matching Algorithms	46
6.2.1	Multi-Way Profile Partitioning	46
6.2.2	A Set Decomposition Problem	50
6.2.2.1	Greedy Clustering	51
6.2.2.2	Giving overlaps more credit	53
6.2.2.3	Clustering: plant the first k seeds	54
6.2.2.4	Incremental Clustering	55
6.2.2.5	Other Heuristics	56
6.3	Updates and Load Re-Balancing	56
7	Automatic Profile Generation	59
7.1	Motivation	59
7.2	Techniques	60
7.3	Applications	60
7.3.1	Active Email System	61
7.3.1.1	Query Generation	62
7.3.1.2	Profile Generation	63
7.3.2	Profile Generation for Categorized Document Collections	67
7.3.2.1	Assumptions	67
7.3.2.2	Ranking Categories	69
7.3.2.3	Subscribing to New Categories	71
7.3.2.4	Unsubscribing from Categories	71

8	Quantitative Evaluation	73
8.1	Methodology	73
8.2	Experiment Setup	74
8.2.1	Platform and Runtime	74
8.2.2	Data Item Collection	75
8.2.3	Profile Generation	75
8.3	Adaptability Measurements	77
8.4	Scalability Measurements	79
8.5	Partitioning Measurements	79
9	Related Work	91
9.1	Historical Development	92
9.2	Content Based Filtering	94
9.3	Social Filtering	96
9.4	State of the Art	98
10	Conclusions	101
10.1	Contributions	102
10.2	Future Work	103
10.2.1	Implementation Extensions	103
10.2.2	Applying Webcasting to New Domains	104
10.3	Final Remarks	105
	Bibliography	107

List of Figures

2.1	Data Collection Architecture	9
2.2	Data Distribution Architecture	10
3.1	Channel Editing Tool	21
4.1	Node Sharing Example	24
4.2	Inserting a Profile	27
4.3	Example for Redundant Internal Nodes	28
4.4	Removing a Profile	29
4.5	Example for Redundant Node (B)	30
4.6	Optimizing the Profile Index	31
4.7	Merging Profile Indexes	32
5.1	Upward and Downward Propagation	35
5.2	Evaluation Process	36
5.3	Downward Propagation	38
5.4	Credit Attribution	40
5.5	Credit Attribution Example	41
5.6	Evaluation with Credit Attribution	42
5.7	Sub-optimal Ranking Example	43
5.8	Propagation for Catalog	43
6.1	Example DAG	48
6.2	Profile Partitions with Duplicate Leaf Nodes	49
6.3	Profile Partitions with References among Leaf Nodes	49
7.1	System Architecture	62
7.2	Patent Classes as Categories with Interest Scores	65
7.3	Example Categories	69

8.1 Adaptability Measurement 80

8.2 Adaptability Measurement (cont.) 81

8.3 Adaptability Measurement (cont.) 82

8.4 Adaptability Measurement (cont.) 83

8.5 Scalability Measurement 84

8.6 Number of Profiles per Partition 86

8.7 Number of Predicates per Partition 87

8.8 Cost per Partition 88

8.9 Matching Performance for Partitions 89

List of Tables

3.1	Reserved Words in Profile Language	16
3.2	Qualifications of Search Words	19
4.1	Attributes of the Profile Index Data Structure	25
8.1	Data Sampling Statistics	76
8.2	Channel Definitions	78