LNCS 3182

Kurt Bauknecht
Martin Bichler
Birgit Pröll (Eds.)

E-Commerce and Web Technologies

5th International Conference, EC-Web 2004
Zaragoza, Spain, August/September 2004
Proceedings

Springer

Kurt Bauknecht Martin Bichler
Birgit Pröll (Eds.)

E-Commerce and Web Technologies

5th International Conference, EC-Web 2004
Zaragoza, Spain, August 31 - September 3, 2004
Proceedings

Springer

Volume Editors

Kurt Bauknecht
University of Zurich, Department of Informatics
Winterthurer Straße 190, CH-8057 Zurich, Switzerland
E-mail: baukn@ifi.unizh.ch

Martin Bichler
Technical University of Munich, Internet-based Information Systems (IBIS)
Boltzmannstr. 3, D-85748 Garching, Germany
E-mail: bichler@in.tum.de

Birgit Pröll
Johannes Kepler University Linz, Institute for Applied Knowledge Processing (FAW)
Altnbergerstr. 69, A-4040 Linz, Austria
E-mail: bproell@faw.uni-linz.ac.at

Library of Congress Control Number: 2004110714

CR Subject Classification (1998): H.4, K.4.4, J.1, K.5, H.3, H.2, K.6.5

ISSN 0302-9743
ISBN 3-540-22917-5 Springer Berlin Heidelberg New York

Springer is a part of Springer Science+Business Media

springeronline.com

Printed in Germany

Typesetting: Camera-ready by author, data conversion by Olgun Computergrafik
Printed on acid-free paper SPIN: 11315261 06/3142 5 4 3 2 1 0

Lecture Notes in Computer Science

For information about Vols. 1–3070

please contact your bookseller or Springer

Vol. 3194: R. Camacho, R. King, A. Srinivasan (Eds.), Inductive Logic Programming. XI, 361 pages. 2004. (Subseries LNAI).

Vol. 3184: S. Katsikas, J. Lopez, G. Pernul (Eds.), Trust and Privacy in Digital Business. XI, 299 pages. 2004.

Vol. 3183: R. Traunmüller (Ed.), Electronic Government. XIX, 583 pages. 2004.

Vol. 3182: K. Bauknecht, M. Bichler, B. Pröll (Eds.), E-Commerce and Web Technologies. XI, 370 pages. 2004.

Vol. 3178: W. Jonker, M. Petkovic (Eds.), Secure Data Management. VIII, 219 pages. 2004.

Vol. 3177: Z.R. Yang, H. Yin, R. Everson (Eds.), Intelligent Data Engineering and Automated Learning – IDEAL 2004. VXIII, 852 pages. 2004.

Vol. 3174: F. Yin, J. Wang, C. Guo (Eds.), Advances in Neural Networks - ISNN 2004. XXXV, 1021 pages. 2004.

Vol. 3172: M. Dorigo, M. Birattari, C. Blum, L. M.Gambardella, F. Mondada, T. Stützle (Eds.), Ant Colony, Optimization and Swarm Intelligence. XII, 434 pages. 2004.

Vol. 3166: M. Rauterberg (Ed.), Entertainment Computing – ICEC 2004. XXIII, 617 pages. 2004.

Vol. 3158: I. Nikolaidis, M. Barbeau, E. Kranakis (Eds.), Ad-Hoc, Mobile, and Wireless Networks. IX, 344 pages. 2004.

Vol. 3157: C. Zhang, H. W. Guesgen, W.K. Yeap (Eds.), PRICAI 2004: Trends in Artificial Intelligence. XX, 1023 pages. 2004. (Subseries LNAI).

Vol. 3156: M. Joye, J.-J. Quisquater (Eds.), Cryptographic Hardware and Embedded Systems - CHES 2004. XIII, 455 pages. 2004.

Vol. 3155: P. Funk, P.A. González Calero (Eds.), Advances in Case-Based Reasoning. XIII, 822 pages. 2004. (Subseries LNAI).

Vol. 3154: R.L. Nord (Ed.), Software Product Lines. XIV, 334 pages. 2004.

Vol. 3153: J. Fiala, V. Koubek, J. Kratochvíl (Eds.), Mathematical Foundations of Computer Science 2004. XIV, 902 pages. 2004.

Vol. 3152: M. Franklin (Ed.), Advances in Cryptology – CRYPTO 2004. XI, 579 pages. 2004.

Vol. 3150: G.-Z. Yang, T. Jiang (Eds.), Medical Imaging and Augmented Reality. XII, 378 pages. 2004.

Vol. 3149: M. Danelutto, M. Vanneschi, D. Laforenza (Eds.), Euro-Par 2004 Parallel Processing. XXXIV, 1081 pages. 2004.

Vol. 3148: R. Giacobazzi (Ed.), Static Analysis. XI, 393 pages. 2004.

Vol. 3146: P. Érdi, A. Esposito, M. Marinaro, S. Scarpetta (Eds.), Computational Neuroscience: Cortical Dynamics. XI, 161 pages. 2004.

Vol. 3144: M. Papatriantafilou, P. Hunel (Eds.), Principles of Distributed Systems. XI, 246 pages. 2004.

Vol. 3143: W. Liu, Y. Shi, Q. Li (Eds.), Advances in Web-Based Learning – ICWL 2004. XIV, 459 pages. 2004.

Vol. 3142: J. Diaz, J. Karhumäki, A. Lepistö, D. Sannella (Eds.), Automata, Languages and Programming. XIX, 1253 pages. 2004.

Vol. 3140: N. Koch, P. Fraternali, M. Wirsing (Eds.), Web Engineering. XXI, 623 pages. 2004.

Vol. 3139: F. Iida, R. Pfeifer, L. Steels, Y. Kuniyoshi (Eds.), Embodied Artificial Intelligence. IX, 331 pages. 2004. (Subseries LNAI).

Vol. 3138: A. Fred, T. Caelli, R.P.W. Duin, A. Campilho, D.d. Ridder (Eds.), Structural, Syntactic, and Statistical Pattern Recognition. XXII, 1168 pages. 2004.

Vol. 3137: P. De Bra, W. Nejdl (Eds.), Adaptive Hypermedia and Adaptive Web-Based Systems. XIV, 442 pages. 2004.

Vol. 3136: F. Meziane, E. Métais (Eds.), Natural Language Processing and Information Systems. XII, 436 pages. 2004.

Vol. 3134: C. Zannier, H. Erdogmus, L. Lindstrom (Eds.), Extreme Programming and Agile Methods - XP/Agile Universe 2004. XIV, 233 pages. 2004.

Vol. 3133: A.D. Pimentel, S. Vassiliadis (Eds.), Computer Systems: Architectures, Modeling, and Simulation. XIII, 562 pages. 2004.

Vol. 3131: V. Torra, Y. Narukawa (Eds.), Modeling Decisions for Artificial Intelligence. XI, 327 pages. 2004. (Subseries LNAI).

Vol. 3130: A. Syropoulos, K. Berry, Y. Haralambous, B. Hughes, S. Peter, J. Plaice (Eds.), TeX, XML, and Digital Typography. VIII, 265 pages. 2004.

Vol. 3129: Q. Li, G. Wang, L. Feng (Eds.), Advances in Web-Age Information Management. XVII, 753 pages. 2004.

Vol. 3128: D. Asonov (Ed.), Querying Databases Privately. IX, 115 pages. 2004.

Vol. 3127: K.E. Wolff, H.D. Pfeiffer, H.S. Delugach (Eds.), Conceptual Structures at Work. XI, 403 pages. 2004. (Subseries LNAI).

Vol. 3126: P. Dini, P. Lorenz, J.N.d. Souza (Eds.), Service Assurance with Partial and Intermittent Resources. XI, 312 pages. 2004.

Vol. 3125: D. Kozen (Ed.), Mathematics of Program Construction. X, 401 pages. 2004.

Vol. 3124: J.N. de Souza, P. Dini, P. Lorenz (Eds.), Telecommunications and Networking - ICT 2004. XXVI, 1390 pages. 2004.

Vol. 3123: A. Belz, R. Evans, P. Piwek (Eds.), Natural Language Generation. X, 219 pages. 2004. (Subseries LNAI).

Vol. 3122: K. Jansen, S. Khanna, J.D.P. Rolim, D. Ron (Eds.), Approximation, Randomization, and Combinatorial Optimization. IX, 428 pages. 2004.

Vol. 3121: S. Nikoletseas, J.D.P. Rolim (Eds.), Algorithmic Aspects of Wireless Sensor Networks. X, 201 pages. 2004.

Vol. 3120: J. Shawe-Taylor, Y. Singer (Eds.), Learning Theory. X, 648 pages. 2004. (Subseries LNAI).

Vol. 3118: K. Miesenberger, J. Klaus, W. Zagler, D. Burger (Eds.), Computer Helping People with Special Needs. XXIII, 1191 pages. 2004.

Vol. 3116: C. Rattray, S. Maharaj, C. Shankland (Eds.), Algebraic Methodology and Software Technology. XI, 569 pages. 2004.

Vol. 3114: R. Alur, D.A. Peled (Eds.), Computer Aided Verification. XII, 536 pages. 2004.

Vol. 3113: J. Karhumäki, H. Maurer, G. Paun, G. Rozenberg (Eds.), Theory Is Forever. X, 283 pages. 2004.

Vol. 3112: H. Williams, L. MacKinnon (Eds.), Key Technologies for Data Management. XII, 265 pages. 2004.

Vol. 3111: T. Hagerup, J. Katajainen (Eds.), Algorithm Theory - SWAT 2004. XI, 506 pages. 2004.

Vol. 3110: A. Juels (Ed.), Financial Cryptography. XI, 281 pages. 2004.

Vol. 3109: S.C. Sahinalp, S. Muthukrishnan, U. Dogrusoz (Eds.), Combinatorial Pattern Matching. XII, 486 pages. 2004.

Vol. 3108: H. Wang, J. Pieprzyk, V. Varadharajan (Eds.), Information Security and Privacy. XII, 494 pages. 2004.

Vol. 3107: J. Bosch, C. Krueger (Eds.), Software Reuse: Methods, Techniques and Tools. XI, 339 pages. 2004.

Vol. 3106: K.-Y. Chwa, J.I. Munro (Eds.), Computing and Combinatorics. XIII, 474 pages. 2004.

Vol. 3105: S. Göbel, U. Spierling, A. Hoffmann, I. Iurgel, O. Schneider, J. Dechau, A. Feix (Eds.), Technologies for Interactive Digital Storytelling and Entertainment. XVI, 304 pages. 2004.

Vol. 3104: R. Kralovic, O. Sykora (Eds.), Structural Information and Communication Complexity. X, 303 pages. 2004.

Vol. 3103: K. Deb, e. al. (Eds.), Genetic and Evolutionary Computation – GECCO 2004. XLIX, 1439 pages. 2004.

Vol. 3102: K. Deb, e. al. (Eds.), Genetic and Evolutionary Computation – GECCO 2004. L, 1445 pages. 2004.

Vol. 3101: M. Masoodian, S. Jones, B. Rogers (Eds.), Computer Human Interaction. XIV, 694 pages. 2004.

Vol. 3100: J.F. Peters, A. Skowron, J.W. Grzymała-Busse, B. Kostek, R.W. Świniarski, M.S. Szczuka (Eds.), Transactions on Rough Sets I. X, 405 pages. 2004.

Vol. 3099: J. Cortadella, W. Reisig (Eds.), Applications and Theory of Petri Nets 2004. XI, 505 pages. 2004.

Vol. 3098: J. Desel, W. Reisig, G. Rozenberg (Eds.), Lectures on Concurrency and Petri Nets. VIII, 849 pages. 2004.

Vol. 3097: D. Basin, M. Rusinowitch (Eds.), Automated Reasoning. XII, 493 pages. 2004. (Subseries LNAI).

Vol. 3096: G. Melnik, H. Holz (Eds.), Advances in Learning Software Organizations. X, 173 pages. 2004.

Vol. 3095: C. Bussler, D. Fensel, M.E. Orlowska, J. Yang (Eds.), Web Services, E-Business, and the Semantic Web. X, 147 pages. 2004.

Vol. 3094: A. Nürnberger, M. Detyniecki (Eds.), Adaptive Multimedia Retrieval. VIII, 229 pages. 2004.

Vol. 3093: S. Katsikas, S. Gritzalis, J. Lopez (Eds.), Public Key Infrastructure. XIII, 380 pages. 2004.

Vol. 3092: J. Eckstein, H. Baumeister (Eds.), Extreme Programming and Agile Processes in Software Engineering. XVI, 358 pages. 2004.

Vol. 3091: V. van Oostrom (Ed.), Rewriting Techniques and Applications. X, 313 pages. 2004.

Vol. 3089: M. Jakobsson, M. Yung, J. Zhou (Eds.), Applied Cryptography and Network Security. XIV, 510 pages. 2004.

Vol. 3087: D. Maltoni, A.K. Jain (Eds.), Biometric Authentication. XIII, 343 pages. 2004.

Vol. 3086: M. Odersky (Ed.), ECOOP 2004 – Object-Oriented Programming. XIII, 611 pages. 2004.

Vol. 3085: S. Berardi, M. Coppo, F. Damiani (Eds.), Types for Proofs and Programs. X, 409 pages. 2004.

Vol. 3084: A. Persson, J. Stirna (Eds.), Advanced Information Systems Engineering. XIV, 596 pages. 2004.

Vol. 3083: W. Emmerich, A.L. Wolf (Eds.), Component Deployment. X, 249 pages. 2004.

Vol. 3080: J. Desel, B. Pernici, M. Weske (Eds.), Business Process Management. X, 307 pages. 2004.

Vol. 3079: Z. Mammeri, P. Lorenz (Eds.), High Speed Networks and Multimedia Communications. XVIII, 1103 pages. 2004.

Vol. 3078: S. Cotin, D.N. Metaxas (Eds.), Medical Simulation. XVI, 296 pages. 2004.

Vol. 3077: F. Roli, J. Kittler, T. Windeatt (Eds.), Multiple Classifier Systems. XII, 386 pages. 2004.

Vol. 3076: D. Buell (Ed.), Algorithmic Number Theory. XI, 451 pages. 2004.

Vol. 3075: W. Lenski (Ed.), Logic versus Approximation. IX, 205 pages. 2004.

Vol. 3074: B. Kuijpers, P. Revesz (Eds.), Constraint Databases and Applications. XII, 181 pages. 2004.

Vol. 3073: H. Chen, R. Moore, D.D. Zeng, J. Leavitt (Eds.), Intelligence and Security Informatics. XV, 536 pages. 2004.

Vol. 3072: D. Zhang, A.K. Jain (Eds.), Biometric Authentication. XVII, 800 pages. 2004.

Vol. 3071: A. Omicini, P. Petta, J. Pitt (Eds.), Engineering Societies in the Agents World. XIII, 409 pages. 2004. (Subseries LNAI).

Preface

We welcome you to the proceedings of the 5t International Conference on E-Commerce and Web Technology (EC-Web 2004) held in conjunction with DEXA 2004 in Zaragoza, Spain. This conference, first held in Greenwich, United Kingdom in 2000, now is in its fifth year and very well established. As in the four previous years, it served as a forum to bring together researchers from academia and commercial developers from industry to discuss the current state of the art in e-commerce and Web technology. Inspirations and new ideas emerged from intensive discussions during formal sessions and social events.

Keynote addresses, research presentations and discussions during the conference helped to further develop the exchange of ideas among the researchers, developers and practitioners present.

The conference attracted 103 paper submissions and almost every paper was reviewed by three program committee members. The program committee selected 37 papers for presentation and publication, a task which was not easy due to the high quality of the submitted papers.

We would like to express our thanks to our colleagues who helped with putting together the technical program: the program committee members and external reviewers for their timely and rigorous reviews of the papers, and the organizing committee for their help in the administrative work and support. We owe special thanks to Gabriela Wagner, Mirella Köster, and Birgit Hauer for their helping hands concerning the administrative and organizational tasks of this conference.

Finally, we would like to thank all the authors who submitted papers, authors who presented papers, and the participants who together made this conference an intellectually stimulating event through their active contributions.

We hope that those who attended enjoyed the hospitality of Zaragoza.

August/September 2004

Martin Bichler
Birgit Pröll

Program Committee

General Chairperson

Kurt Bauknecht, University of Zurich, Switzerland

Conference Program Chairpersons

Martin Bichler, Technical University of Munich, Germany
Birgit Proell, FAW, Johannes Kepler University, Linz, Austria

Program Committee Members

Sourav S. Bhowmick, Nanyang Technological University, Singapore
Susanne Boll, University of Oldenburg, Germany
Walter Brenner, University of St. Gallen, Switzerland
Stephane Bressan, National University of Singapore, Singapore
Tanya Castleman, Deakin University, Australia
Wojciech Cellary, Poznan University of Economics, Poland
Jen-Yao Chung, IBM T.J. Watson Research Center, USA
Roger Clarke, The Australian National University, Australia
Asuman Dogac, Middle East Technical University, Turkey
Eduardo Fernandez, Florida Atlantic University, USA
Elena Ferrari, University of Insubria at Como, Italy
Farshad Fotouhi, Wayne State University, USA
Yongjian Fu, Cleveland State University, USA
Chanan Glezer, Ben Gurion University of the Negev, Israel
Rüdiger Grimm, University of Technology, Ilmenau, Germany
Manfred Hauswirth, EPFL, Switzerland
Thomas Hess, LMU Munich, Germany
Yigal Hoffner, IBM Zurich Research Laboratory, Switzerland
Christian Huemer, University of Vienna, Austria
Gregory E. Kersten, University of Ottawa, Canada
Hiroyuki Kitagawa, University of Tsukuba, Japan
Wolfgang Koenig, Frankfurt University, Germany
Gabriele Kotsis, Johannes Kepler University, Linz, Austria
Winfried Lamersdorf, University of Hamburg, Germany
Alberto Laender, Federal University of Minas Gerais, Brazil
Juhnyoung Lee, IBM T.J. Watson Research Center, USA
Leszek Lilien, Purdue University, USA
Ee-Peng Lim, Nanyang Technological University, Singapore
Huan Liu, Arizona State University, USA

Heiko Ludwig, IBM T.J. Watson Research Center, USA
Sanjay Kumar Madria, University of Missouri-Rolla, USA
Bamshad Mobasher, DePaul University, USA
Natwar Modani, Oracle, India
Mukesh Mohania, IBM India Research Lab, India
Gustaf Neumann, Vienna University of Economics and Business Administration, Austria
Wee Keong Ng, Nanyang Technological University, Singapore
Rolf Oppliger, eSECURITY Technologies, Switzerland
Stefano Paraboschi, University of Bergamo, Italy
Oscar Pastor, Valencia University of Technology, Spain
Günther Pernul, University of Regensburg, Germany
Evangelia Pitoura, University of Ioannina, Greece
Gerald Quirchmayr, University of Vienna, Austria
Indrakshi Ray, Colorado State University, USA
Werner Retschitzegger, Johannes Kepler University, Linz, Austria
Tomas Sabol, Technical University of Kosice, Slovakia
Nandlal L. Sarda, Indian Institute of Technology, Bombay, India
Michael Stroebel, BMW Group, Germany
Roger M. Tagg, University of South Australia, Australia
Kian-Lee Tan, National University of Singapore, Singapore
Stephanie Teufel, University of Fribourg, Switzerland
Bruce H. Thomas, University of South Australia, Australia
A Min Tjoa, Technical University of Vienna, Austria
Aphrodite Tsalgatidou, National and Kapodistrian University of Athens, Greece
Krishnamurthy Vidyasankar, Memorial Univ., St. John's, Newfoundland, Canada
Hans Weigand, Tilburg University, The Netherlands
Christof Weinhardt, University of Karlsruhe, Germany
Hannes Werthner, University of Trento, Italy
Andrew Whinston, University of Texas, USA

External Reviewers

Bugrahan Akcay, Middle East Technical University, Turkey
George Athanasopoulos, University of Athens, Greece
Peter Bednar, Technical University of Kosice, Slovakia
Daniel Beimborn, Frankfurt University, Germany
Stefan Blumenberg, Frankfurt University, Germany
Sudip Chakraborty, Colorado State University, USA
Radoslav Delina, Technical University of Kosice, Slovakia
Rainer B. Fladung, Frankfurt University, Germany
Ozgur Gulderen, Middle East Technical University, Turkey
Yavuz Gurcan, Middle East Technical University, Turkey
Yildiray Kabak, Middle East Technical University, Turkey
George-Dimitrios Kapos, Harokopio Univ. of Athens, Greece

Eunjong Kim, Colorado State University, USA
Eleni Koutrouli, University of Athens, Greece
Gokce Banu Laleci, Middle East Technical University, Turkey
Marian Mach, Technical University of Kosice, Slovakia
Sebastian F. Martin, Frankfurt University, Germany
Jan Paralic, Technical University of Kosice, Slovakia
Willy Picard, Poznan University of Economics, Poland
Thomi Pilioura, University of Athens, Greece
Indrajit Ray, Colorado State University, USA
Veshaal Singh, Oracle India Pvt. Ltd., India
Siyamed Sinir, Middle East Technical University, Turkey
Seda Unal, Middle East Technical University, Turkey
Abhishek Verma, Oracle India Pvt. Ltd., India
Murat Yukselen, Middle East Technical University, Turkey

Table of Contents

Recommender Systems

Databases and EC Applications

Service-Oriented e-Commerce Applications

Electronic Negotiation Systems

Security and Trust in e-Commerce

Techniques for B2B e-Commerce

Negotiation Strategies and Protocols

Modeling of e-Commerce Applications

Using Attributes to Improve Prediction Quality in Collaborative Filtering

Taek-Hun Kim and Sung-Bong Yang

Dept. of Computer Science, Yonsei University
Seoul, 120-749, Korea
{kimthun,yang}@cs.yonsei.ac.kr

Abstract. To save customers' time and efforts in searching the goods in the Internet, a customized recommender system is required. It is very important for a recommender system to predict accurately by analyzing customer's preferences. A recommender system utilizes in general an information filtering technique called collaborative filtering, which is based on the ratings of other customers who have similar preferences. Because a recommender system using collaborative filtering predicts customer's preferences based only on the items without useful information on the attributes of each item, it may not give high quality recommendation consistently to the customers. In this paper we show that exploiting the attributes of each item improves prediction quality. We analyze the dataset and retrieve the preferences for the attributes because they have not been rated by customers explicitly. In the experiment the MovieLens dataset of the GroupLens Research Center has been used. The results on various experiments using several neighbor selection methods which are quite popular techniques for recommender systems show that the recommender systems using the attributes provide better prediction qualities than the systems without using the attribute information. Each of the systems using the attributes has improved the prediction quality more than 9%, compared with its counterpart. And the clustering-based recommender systems using the attributes can solve the very large-scale dataset problem without deteriorating prediction quality.

1 Introduction

Nowadays, customers spend so much time and efforts in finding the best suitable goods since more and more information is placed on-line. To save their time and efforts in searching the goods they want, a customized recommender system is required. A customized recommender system should predict the most suitable goods for customers by retrieving and analyzing their preferences. A recommender system utilizes in general an information filtering technique called collaborative filtering which is widely used for such recommender systems as Amazon.com and CDNow.com[1][2][3][4]. A recommender system using collaborative filtering, we call it *CF*, is based on the ratings of the customers who have similar preferences with respect to a given (test) customer.

K. Bauknecht, M. Bichler, and B. Pröll (Eds.): EC-Web 2004, LNCS 3182, pp. 1–10, 2004.

CF calculates the similarity between the test customer and each of other customers who have rated the items that are already rated by the test customer. CF uses in general the Pearson correlation coefficient for calculating the similarity, but it assumes that there must exist at least one item which has already been rated by both the test customer and one of the other customers. A weak point of the "pure" CF is that it uses all other customers including "useless" customers as the neighbors of the test customer. Since CF is based on the ratings of the neighbors who have similar preferences, it is very important to select the neighbors properly to improve prediction quality. Another weak point of CF is that it never considers customer's preferences on the attributes of each item.

There have been many investigations to select proper neighbors based on neighbor selection methods such as the k-nearest neighbor selection, the threshold-based neighbor selection, and the clustering-based neighbor selection. They are quite popular techniques for recommender systems[1][4][6][7][10]. These techniques then predict customer's preferences for the items based on the results of the neighbors' evaluation on the same items.

The recommender system with the clustering-based neighbor selection method is known to predict with worse accuracy than the systems with other neighbor selection methods, but it can solve the very large-scale problem in recommender systems. With millions of customers and items, a recommender system running existing algorithms will suffer serious scalability problem. So it is needed a new idea that can quickly produce high prediction quality, even for the very large-scale problem[8][10].

CF works quite well in general, because it is based on the ratings of other customers who have similar preferences. However, CF may not provide high quality recommendations for the test customer consistently, because it does not consider the attributes of each item and because it depends only on the ratings of other customers who rated the items that are already rated by the test customer. To improve prediction quality, CF needs reinforcements such as utilizing "useful" attributes of the items and using a more refined neighbor selection method. Item attributes mean its originality of the item it holds which differs from the others. In general, they can be obtained through the properties of the goods in real world.

In this paper we show that exploiting the attributes of each item improves prediction quality. We analyze the dataset and retrieve the preferences for the attributes, because they have not been rated by the customers explicitly. In the experiment the MovieLens dataset of the GroupLens Research Center has been used[11]. The dataset consists of 100,000 preferences for 1,682 movies rated by 943 customers explicitly.

We show the impact of utilizing the attributes information on the prediction qualities through various experiments. The experimental results show that the recommender systems using the attribute information provide better prediction qualities than other methods that do not utilize the attributes. Each of the systems using the attributes has improved the prediction quality more than 9%, compared with its counterpart. And besides the clustering-based CF using the attributes can solve the very large-scale problem without deteriorating prediction quality.

The rest of this paper is organized as follows. Section 2 describes briefly CF and several neighbor selection methods. Section 3 illustrates how to utilize the attributes of items for recommender systems in detail. In Section 4, the experimental results are presented. Finally, the conclusions are given in Section 5.

2 Collaborative Filtering and Neighbor Selection Methods

CF recommends items through building the profiles of the customers from their preferences for each item. In CF, preferences are represented generally as numeric values which are rated by the customers. Predicting the preference for a certain item that is new to the test customer is based on the ratings of other customers for the 'target' item. Therefore, it is very important to find a set of customers, called *neighbors*, with more similar preferences to the test customer for better prediction quality.

In CF, Equation (1) is used to predict the preference of a customer. Note that in the following equation $w_{a,k}$ is the Pearson correlation coefficient as in Equation (2) [2][3][4][6][8].

$$P_{a,i} = \overline{r_a} + \frac{\sum_k \{w_{a,k} \times (r_{k,i} - \overline{r_k})\}}{\sum_k \mid w_{a,k} \mid} \,. \tag{1}$$

$$w_{a,k} = \frac{\sum_j (r_{a,j} - \overline{r_a})(r_{k,j} - \overline{r_k})}{\sqrt{\sum_j (r_{a,j} - \overline{r_a})^2 \sum_j (r_{k,j} - \overline{r_k})^2}} \,. \tag{2}$$

In the above equations $P_{a,i}$ is the preference of customer a with respect to item i. $\overline{r_a}$ and $\overline{r_k}$ are the averages of customer a's ratings and customer k's ratings, respectively. $r_{k,i}$ and $r_{k,j}$ are customer k's ratings for items i and j, respectively, and $r_{a,j}$ is customer a's rating for item j.

If customers a and k have similar ratings for an item, $w_{a,k} > 0$. We denote $\mid w_{a,k} \mid$ to indicate how much customer a tends to agree with customer k on the items that both customers have already rated. In this case, customer a is a "positive" neighbor with respect to customer k, and vice versa. If they have opposite ratings for an item, then $w_{a,k} < 0$. Similarly, customer a is a "negative" neighbor with respect to customer k, and vice versa. In this case $\mid w_{a,k} \mid$ indicates how much they tend to disagree on the item that both again have already rated. Hence, if they don't correlate each other, then $w_{a,k} = 0$. Note that $w_{a,k}$ can be in between -1 and 1 inclusive.

Although CF can be regarded as a good choice for a recommender system, there is still much more room for improvement in prediction quality. To do so, CF needs reinforcements such as utilizing "useful" attributes of the items as well as a more refined neighbor selection. In the rest of this section we describe several neighbor selection methods.

2.1 The k-Nearest Neighbor Selection

The k-nearest neighbor method selects the nearest k neighbors who have similar preferences to the test customer by computing the similarities based on their preferences. It only uses the k neighbors who have higher correlation with the test customer than others, while CF suffers from considering all the customers in the input dataset for calculating the preference of the test customer. Thus CF should even consider some customers who may give bad influences on prediction quality. It has been shown in several investigations that the recommender system with the k-nearest neighbor selection method has better quality of prediction than the "pure" CF[1][4][6].

2.2 The Threshold-Based Neighbor Selection

The threshold-based neighbor selection method selects the neighbors who belong to a certain range with respect to the similarities of the preferences. Contrary to the k-nearest neighbor selection method, the number of neighbors selected by this method varies, because it selects neighbors according to a certain threshold value τ.

In the recommender system with the threshold-based neighbor selection, the positive neighbors whose correlations to the test customer are greater than and equal to τ are selected as the neighbors[6]. However, it is also needed that we include the "negative" neighbors whose correlations to the test customer are less than and equal to τ, because they could contribute toward the better neighbor selection for the test customer as they provide negative "opinions" to the test customer. It is obvious that selecting only negative neighbors results in worse prediction qualities than the case in which only positive neighbors are selected, intuitively.

2.3 The Clustering-Based Neighbor Selection

The k-means clustering method creates k clusters each of which consists of the customers who have similar preferences among themselves. In this method we first select k customers arbitrarily as the initial center points of the k clusters, respectively. Then each customer is assigned to a cluster in such a way that the distance between the customer and the center of a cluster is minimized. The distance is calculated using the Euclidean distance, that is, a square root of the element-wise square of the difference between the customer and each center point. The Pearson correlation coefficient can be substituted for the Euclidean distance.

We then calculate the mean of each cluster based on the customers who currently belong to the cluster. The mean is now considered as the new center of