Encyclopedia of Data Warehousing and Mining

Second Edition

John Wang
Montclair State University, USA

Volume I
A–Data Pre

Information Science REFERENCE

INFORMATION SCIENCE REFERENCE
Hershey · New York

Director of Editorial Content: Kristin Klinger
Director of Production: Jennifer Neidig
Managing Editor: Jamie Snavely
Assistant Managing Editor: Carole Coulson
Typesetter: Amanda Appicello, Jeff Ash, Mike Brehem, Carole Coulson, Elizabeth Duke, Jen Henderson, Chris Hrobak, Jennifer Neidig, Jamie Snavely, Sean Woznicki
Cover Design: Lisa Tosheff
Printed at: Yurchak Printing Inc.

Published in the United States of America by
Information Science Reference (an imprint of IGI Global)
701 E. Chocolate Avenue, Suite 200
Hershey PA 17033
Tel: 717-533-8845
Fax: 717-533-8661
E-mail: cust@igi-global.com
Web site: http://www.igi-global.com/reference

and in the United Kingdom by
Information Science Reference (an imprint of IGI Global)
3 Henrietta Street
Covent Garden
London WC2E 8LU
Tel: 44 20 7240 0856
Fax: 44 20 7379 0609
Web site: http://www.eurospanbookstore.com

Library of Congress Cataloging-in-Publication Data

Encyclopedia of data warehousing and mining / John Wang, editor. -- 2nd ed.
p. cm.
Includes bibliographical references and index.
Summary: "This set offers thorough examination of the issues of importance in the rapidly changing field of data warehousing and mining"--Provided by publisher.
ISBN 978-1-60566-010-3 (hardcover) -- ISBN 978-1-60566-011-0 (ebook)
1. Data mining. 2. Data warehousing. I. Wang, John,
QA76.9.D37E52 2008
005.74--dc22
2008030801

British Cataloguing in Publication Data
A Cataloguing in Publication record for this book is available from the British Library.

Bridging Taxonomic Semantics to Accurate Hierarchical Classification

Lei Tang
Arizona State University, USA

Huan Liu
Arizona State University, USA

Jiangping Zhang
The MITRE Corporation, USA

INTRODUCTION

The unregulated and open nature of the Internet and the explosive growth of the Web create a pressing need to provide various services for content categorization. The hierarchical classification attempts to achieve both accurate classification and increased comprehensibility. It has also been shown in literature that hierarchical models outperform flat models in training efficiency, classification efficiency, and classification accuracy (Koller & Sahami, 1997; McCallum, Rosenfeld, Mitchell & Ng, 1998; Ruiz & Srinivasan ,1999; Dumais & Chen, 2000; Yang, Zhang & Kisiel, 2003; Cai & Hofmann, 2004; Liu, Yang, Wan, Zeng, Cheng & Ma, 2005). However, the quality of the taxonomy attracted little attention in past works. Actually, different taxonomies can result in differences in classification. So the quality of the taxonomy should be considered for real-world classifications. Even a semantically sound taxonomy does not necessarily lead to the intended classification performance (Tang, Zhang & Liu 2006). Therefore, it is desirable to construct or modify a hierarchy to better suit the hierarchical content classification task.

BACKGROUND

Hierarchical models rely on certain predefined content taxonomies. Content taxonomies are usually created for ease of content management or access, so semantically similar categories are grouped into a parent category. Usually, a subject expert or librarian is employed to organize the category labels into a hierarchy using some ontology information. However, such a taxonomy is often generated independent of data (e.g., documents). Hence, there may exist some inconsistency between the given taxonomy and data, leading to poor classification performance.

First, semantically similar categories may not be similar in lexical terms. Most content categorization algorithms are statistical algorithms based on the occurrences of lexical terms in content. Hence, a semantically sound hierarchy does not necessarily lead to the intended categorization result.

Second, even for the same set of categories, there could be different semantically sound taxonomies. Semantics does not guarantee a unique taxonomy. Different applications may need different category taxonomies. For example, sports teams may be grouped according to their locations such as Arizona, California, Oregon, etc and then the sports types such as football, basketball, etc.. Depending upon the application, they may also be grouped according to the sports types first and then locations. Both taxonomies are reasonable in terms of semantics. With a hierarchical classification model, however, the two taxonomies would likely result in different performances. Hence, we need to investigate the impact of different hierarchies (taxonomies) on classification.

In addition, semantics may change over time. For example, when the semantic taxonomy was first generated, people would not expect the category *Hurricane* related to *Politics*, and likely put it under *Geography*. However, after investigating the data recently collected, it is noticed that a good number of documents in category *Hurricane* are actually talking about the disasters Hurricane Katrina and Rita in the United States and the responsibility and the faults of FEMA during the crises. Based on the content, it is more reasonable to put

Hurricane under *Politics* for better classification. This example demonstrates the stagnant nature of taxonomy and the dynamic change of semantics reflected in data. It also motivates the data-driven adaptation of a given taxonomy in hierarchical classification.

MAIN FOCUS

In practice, semantics based taxonomies are always exploited for hierarchical classification. As the taxonomic semantics might not be compatible with specific data and applications and can be ambiguous in certain cases, the semantic taxonomy might lead hierarchical classifications astray. There are mainly two directions to obtain a taxonomy from which a good hierarchical model can be derived: *taxonomy generation via clustering* or *taxonomy adaptation via classification learning*.

Taxonomy Generation via Clustering

Some researchers propose to generate taxonomies from data for document management or classification. Note that the taxonomy generated here focus more on comprehensibility and accurate classification, rather than efficient storage and retrieval. Therefore, we omit the tree-type based index structures for high-dimensional data like R*-tree (Beckmann, Kriegel, Schneider & Seeger 1990), TV-tree (Lin, Jagadish & Faloutsos 1994), etc. Some researchers try to build a taxonomy with the aid of human experts (Zhang, Liu, Pan & Yang 2004, Gates, Teiken & Cheng 2005) whereas other works exploit some hierarchical clustering algorithms to automatically fulfill this task. Basically, there are two approaches for hierarchical clustering*: agglomerative* and *divisive*.

In Aggarwal, Gates & Yu (1999), Chuang & Chien (2004) and Li & Zhu (2005), all employ a hierarchical *agglomerative* clustering (HAC) approach. In Aggarwal, Gates & Yu (1999), the centroids of each class are used as the initial seeds and then projected clustering method is applied to build the hierarchy. During the process, a cluster with few documents is discarded. Thus, the taxonomy generated by this method may have different categories than predefined. The authors evaluated their generated taxonomies by some user study and found its perfomrance is comparable to the Yahoo directory. In Li & Zhu (2005), a linear discriminant projection is applied to the data first and then a hierarchical clustering method UPGMA (Jain & Dubes 1988) is exploited to generate a dendrogram which is a binary tree. For classification, the authors change the dendrogram to a two-level tree according to the cluster coherence, and hierarchical models yield classification improvement over flat models. But it is not sufficiently justified why a two-level tree should be adopted. Meanwhile, a similar approach, HAC+P was proposed by Chuang & Chien (2004). This approach adds one post-processing step to automatically change the binary tree obtained from HAC, to a wide tree with multiple children. However, in this process, some parameters have to be specified as the maximum depth of the tree, the minimum size of a cluster, and the cluster number preference at each level. These parameters make this approach rather ad hoc.

Comparatively, the work in Punera, Rajan & Ghosh (2005) falls into the category of *divisive* hierarchical clustering. The authors generate a taxonomy in which each node is associated with a list of categories. Each leaf node has only one category. This algorithm basically uses the centroids of the two most distant categories as the initial seeds and then applies Spherical K-Means (Dhillon, Mallela & Kumar, 2001) with k=2 to divide the cluster into 2 sub-clusters. Each category is assigned to one sub-cluster if majority of its documents belong to the sub-cluster (its ratio exceeds a predefined parameter). Otherwise, this category is associated to both sub-clusters. Another difference of this method from other HAC methods is that it generates a taxonomy with one category possibly occurring in multiple leaf nodes.

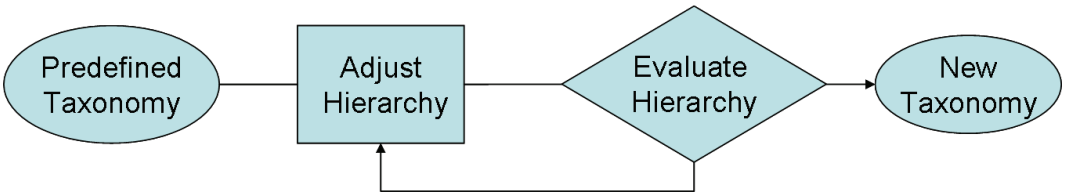
Taxonomy Adaptation via Classification Learning

Taxonomy clustering approach is appropriate if no taxonomy is provided at the initial stage. However, in reality, a human-provided semantic taxonomy is almost always available. Rather than "start from scratch", Tang, Zhang & Liu (2006) proposes to adapt the predefined taxonomy according the classification result on the data.

Three elementary hierarchy adjusting operations are defined:

- **Promote:** Roll up one node to upper level;
- **Demote:** Push down one node to its sibling;

Figure 1.



- **Merge:** Merge two sibling nodes to form a super node; Then a wrapper approach is exploited as in seen Figure 1.

The basic idea is, given a predefined taxonomy and training data, a different hierarchy can be obtained by performing the three elementary operations. Then, the newly generated hierarchy is evaluated on some validation data. If the change results in a performance improvement, we keep the change; otherwise, new change to the original taxonomy is explored. Finally, if no more change can lead to performance improvement, we output the new taxonomy which acclimatizes the taxonomic semantics according to the data.

In (Tang, Zhang & Liu 2006), the hierarchy adjustment follows a top-down traversal of the hierarchy. In the first iteration, only promoting is exploited to adjust the hierarchy whereas in the next iteration, demoting and merging are employed. This pair-wise iteration keeps running until no performance improvement is observed on the training data. As shown in their experiment, two iterations are often sufficient to achieve a robust taxonomy for classification which outperforms the predefined taxonomy and the taxonomy generated via clustering.

FUTURE TRENDS

Taxonomy can be considered as a form of prior knowledge. Adapting the prior knowledge to better suit the data is promising and desirable. Current works either abandon the hierarchy information or start taxonomy adaptation using a wrapper model. This short article provides some starting points that can hopefully lead to more effective and efficient methods to explore the prior knowledge in the future. When we better understand the problem of hierarchical classification and hierarchy consistency with data, we will investigate how to provide a filter approach which is more efficient to accomplish taxonomy adaptation.

This problem is naturally connected to Bayesian inference as well. The predefined hierarchy is the prior and the newly generated taxonomy is a "posterior" hierarchy. Integrating these two different fields—data mining and Bayesian inference, to reinforce the theory of taxonomy adaptation and to provide effective solution is a big challenge for data mining practitioners.

It is noticed that the number of features selected at each node can affect the performance and the structure of a hierarchy. When the class distribution is imbalanced, which is common in real-world applications, we should also pay attention to the problem of feature selection in order to avoid the bias associated with skewed class distribution (Forman, 2003; Tang & Liu, 2005). An effective criterion to select features can be explored in combination with the hierarchy information in this regard. Some general discussions and research issues of feature selection can be found in Liu & Motoda (1998) and Liu & Yu (2005).

CONCLUSION

Hierarchical models are effective for classification when we have a predefined semantically sound taxonomy. Since a given taxonomy may not necessarily lead to the best classification performance. Our task is how to obtain a data-driven hierarchy so that a reasonably good classifier can be inducted. In this article, we present an initial attempt to review and categorize the existing approaches: *taxonomy generation via clustering* and *taxonomy adaptation via classification learning*. It is anticipated that this active area of research will produce more effective and efficient approaches that are likely to emerge in a vast range of applications of web mining and text categorization.

REFERENCES

Aggarwal, C.C., Gates, S.C. & Yu, P.S. (1999). On the merits of building categorization systems by supervised clustering. *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 281-282).

Beckmann, N., Kriegel, H., Schneider, R. & Seeger, B.(1990) The R*-tree: an efficient and robust access method for points and rectangels. *Proceedings of the ACM SIGMOD international conference on management of data* (pp.322-331).

Cai, L. & Hofmann, T. (2004). Hierarchical document categorization with support vector machines, *Proceedings of the thirteenth ACM conference on Information and knowledge management* (pp. 78-87).

Chuang, S. & Chien, L. (2004). A practical web-based approach to generating topic hierarchy for text segments, *Proceedings of the thirteenth ACM conference on Information and knowledge management* (pp. 127-136)

Dhillon, I.S., Mallela, S. & Kumar, R. (2001) Efficient Clustering of Very Large Document Collections, in *Data Mining for Scientific and Engineering Applications*, Kluwer Academic.

Dumais, S. & Chen, H. (2000). Hierarchical classification of Web content. *Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 256-263).

Forman, G. (2003). An Extensive Empirical Study of Feature Selection Metrics for Text Classification. *Journal of Machine Learning Research*, 3, 1289-1305.

Gates, S. C., Teiken, W., and Cheng, K. F. (2005). Taxonomies by the numbers: building high-performance taxonomies. In *Proceedings of the 14th ACM international Conference on information and Knowledge Management*. (pp. 568-577).

Jain, A.K.& Dubes, R.C. (Ed.). (1988). *Algorithms for clustering data.* Prentice-Hal Inc

Koller, D. & Sahami, M.(1997). Hierarchically classifying documents using very few words, *Proceedings of the Fourteenth International Conference on Machine Learning* (pp. 170-178).

Li, T. & Zhu, S. (2005). Hierarchical document classification using automatically generated hierarchy. *Proceedings of SIAM 2005 Data Mining Conference*(pp. 521-525).

Lin, K. I., Jagadish, H. V., and Faloutsos, C. 1994. The TV-tree: an index structure for high-dimensional data. *The VLDB Journal* 3, 4 (Oct. 1994), 517-542.

Liu, H. & Motoda, H.(Ed.). (1998). *Feature selection for knowledge discovery and data mining* Boston: Kluwer Academic Publishers.

Liu, H. & Yu, L. (2005). Toward integrating feature selection algorithms for classification and clustering. *IEEE Transactions on Knowledge and Data Engineering*, 17, 491-502.

Liu, T., Yang, Y., Wan, H., Zeng, H., Chen, Z. & Ma, W. (2005). Support vector machines classification with a very large-scale taxonomy, *SIGKDD Explor. Newsl.*, 7(1), 36-43.

McCallum, A., Rosenfeld, R., Mitchell, T.M. & Ng, A.Y. (1998). Improving text classification by shrinkage in a hierarchy of classes, *Proceedings of the Fifteenth International Conference on Machine Learning* (pp. 359-367).

Punera, K., Rajan, S. & Ghosh, J. (2005). Automatically learning document taxonomies for hierarchical classification, *Special interest Tracks and Posters of the 14th international Conference on World Wide Web*(pp. 1010-1011).

Ruiz, M.E. & Srinivasan, P. (1999). Hierarchical neural networks for text categorization, *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 281-282).

Tang, L. & Liu, H. (2005) Bias analysis in text classification for highly skewed data, *Proceedings of the 5th IEEE international conference on Data Mining* (pp. 781-784).

Tang, L., Zhang, J. & Liu, H. (2006) *Acclimatizing taxonomic semantics for hierarchical content categorization Proceedings of the 12th Annual SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 384-393)

Yang, Y., Zhang, J. & Kisiel, B. (2003) A scalability analysis of classifiers in text categorization, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval* (pp. 96-103).

Zhang, L., Liu, S., Pan, Y., and Yang, L. 2004. InfoAnalyzer: a computer-aided tool for building enterprise taxonomies. In *Proceedings of the Thirteenth ACM international Conference on information and Knowledge Management* (pp. 477-483).

KEY TERMS

Classification: A process of predicting the classes of unseen instances based on patterns learned from available instances with predefined classes.

Clustering: A process of grouping instances into clusters so that instances are similar to one another within a cluster but dissimilar to instances in other clusters.

Filter Model: A process that selects the best hierarchy without building hierarchical models.

Flat Model: A classifier outputs a classification from the input without any intermediate steps.

Hierarchical Model: A classifier outputs a classification using the taxonomy information in intermediate steps.

Hierarchy/Taxonomy: A tree with each node representing a category. Each leaf node represents a class label we are interested.

Taxonomy Adaptation: A process which adapts the predefined taxonomy based on the some data with class labels. All the class labels appear in the leaf node of the newly generated taxonomy.

Taxonomy Generation: A process which generates taxonomy based on some data with class labels so that all the labels appear in the leaf node of the taxonomy.

Wrapper Model: A process which builds a hierarchical model on training data and evaluates the model on validation data to select the best hierarchy. Usually, this process involves multiple constructions and evaluations of hierarchical models.