

Multi-labeled data expressed by a set of labels

Furukawa, Tetsuya
Department of Economic Engineering, Kyushu University

Kuzunishi, Masahiro
Faculty of Business and Commerce

<https://hdl.handle.net/2324/26640>

出版情報 : World Academy of Science, Engineering and Technology. 41, pp.577-583, 2010-05-01.
World Academy of Science, Engineering and Technology (WASET)

バージョン :

権利関係 : (C) 2010 World Academy of Science, Engineering and Technology

Multi-labeled Data Expressed by a Set of Labels

Tetsuya Furukawa and Masahiro Kuzunishi

Abstract—Collected data must be organized to be utilized efficiently, and hierarchical classification of data is efficient approach to organize data. When data is classified to multiple categories or annotated with a set of labels, users request multi-labeled data by giving a set of labels. There are several interpretations of the data expressed by a set of labels. This paper discusses which data is expressed by a set of labels by introducing orders for sets of labels and shows that there are four types of orders, which are characterized by whether the labels of expressed data includes every label of the given set of labels within the range of the set. Desirable properties of the orders, data is also expressed by the higher set of labels and different sets of labels express different data, are discussed for the orders.

Keywords—Classification Hierarchies, Multi-labeled Data, Multiple Classification, Orders of Sets of Labels

I. INTRODUCTION

PROGRESS of information technologies and arrangement of network environments have been increasing available data including various kinds such as numerical data, texts, images, audio, etc. With the remarkable growth of data, it is becoming increasingly important to organize collected data properly. Hierarchical classification based on the content of data is one of the efficient methods to organize such data [2] [10] [11], which is used in the category searches in search engines, for example. Data is classified to categories or annotated with the labels of the categories.

Data is usually assumed to be classified to one category, which is called single-label classification [2] [14]. In *News-groups* data set, each news document is classified to only one category [12]. However, there is data which should be classified to multiple categories. For example, data on a comparison between manufacturing and financial industries should not be classified to either category *Manufacture* or *Finance* but to both in the classification for an industrial type. Such data is classified with multi-label classification, where data is classified to multiple categories [1] [8] [12]. In multi-label classification, the data on a comparison between manufacturing and financial industries is classified to both categories *Manufacture* and *Finance*, and labeled {*Manufacture*, *Finance*}.

Users or applications request data by giving labels. There are two kinds of “data identified by a label,” the data with the same label as the given label and the data with a label whose concept is lower than or equal to the concept of the given label [7]. The data identified by label *Manufacture* is

the data labeled *Manufacture* and the data with one of labels *Manufacture*, *Transportation*, *Automobile*, etc., respectively. In utilization of classified data, the latter is usually adopted, which this paper focuses on. When data is classified with single-label classification, the utilization of the data is rather straightforward. In multi-label classification, a set of labels can be used to identify a set of multi-labeled data because data have multiple labels. There are several kinds of “data identified by a set of labels.”

Example 1 Suppose a set of labels $L = \{\textit{Manufacture}, \textit{Finance}\}$. The data identified by L is usually regarded as “the data related to nothing but manufacturing and financial industries” such as data labeled {*Automobile*, *Credit*}. On the other hand, there can be other sets of data identified by L . When the data identified by L means “the data related to manufacturing and financial industries,” it includes data labeled {*Automobile*, *Credit*, *Medicine*} where *Medicine* is not related to *Manufacture* or *Finance*. There are also such meanings that “the data related to only manufacturing industry or finance industry” and “the data related to manufacturing industry or finance industry,” which include data labeled {*Automobile*} and {*Automobile*, *Medicine*}, with no label for *Finance*, respectively. \square

Although there are several kinds of data identified by a set of labels, there is few discussions on the semantics shown in Example 1. Recent researches on classification allow multi-labeled data such as Web and texts [6] [11], whose purpose is automatic classification of data to multiple categories, and data is used through intersection or union of categories. In the utilization of multi-labeled data, methods to find the data matching given set of keywords are developed [3] [4], which rank data by frequency of keywords and their relationships so that users can find data satisfying their criteria. In those researches, the data identified by a given set of labels are such data as “the data related to all of the labels” or “the data related to any of the labels.”

To utilize multi-labeled data precisely, there must be advanced usage based on the multiple labels. This paper introduces orders for sets of labels so that data is expressed by a set of labels if the label of the data is lower than or equal to the set of labels. Data is identified by a set of labels as the data expressed by the set of labels.

Usually a set of labels is interpreted as conjunction or disjunction of the elements, that is, the intersection or the union of the sets of data for the labels. These bring two types of orders for sets of labels. Other orders also exist, and those orders for sets of labels appear by systematic discussion. The purpose of this paper is to formalize the various possible orders.

T. Furukawa is Professor of Dept. of Economic Engineering, Kyushu University, Hakozaki 6-19-1, Higashi-ku, Fukuoka 812-8581 Japan (e-mail: furukawa@en.kyushu-u.ac.jp.)

M. Kuzunishi is Assistant Professor of Faculty of Business and Commerce, Aichi Gakuin University, Araiike 12, Iwasaki, Nisshin, Aichi 470-0195 Japan (e-mail: kuzunishi@dpc.agu.ac.jp)

of multi-labeled data [1] [8] [12]. The results of this paper can be applied to such fields.

REFERENCES

- [1] G. Adami, P. Avesani, and D. Sona, "Bootstrapping for Hierarchical Document Classification," *Proc. Int'l Conf. on Information and Knowledge Management (CIKM'03)*, pp. 295–302, 2003.
- [2] E. Bertino, J. Fan, E. Ferrari, M. Hachi, and A. Elamagarmid, "A Hierarchical Access Control Model for Video Database Systems," *ACM Transactions on Information Systems*, Vol.21, No.2, pp. 151–191, 2003.
- [3] K. Chakrabarti, V. Ganti, J. Han, and D. Xin, "Ranking Objects by Exploiting Relationships: Computing Top-K over Aggregation," *Proc. ACM SIGMOD Int'l Conf. on Management of Data*, pp. 371–382, 2006.
- [4] S. Chuang and L. Chien, "Taxonomy Generation for Text Segments: A Practical Web-Based Approach," *ACM Transactions on Information Systems*, Vol.23, No.4, pp. 363–396, 2005.
- [5] W. Dakka, P. G. Ipeirotis, and K. R. Wood, "Automatic Construction of Multifaceted Browsing Interfaces," *Proc. Int'l Conf. on Information and Knowledge Management (CIKM'05)*, pp. 768–775, 2005.
- [6] S. Dumais and H. Chen, "Hierarchical Classification of Web Content," *Proc. ACM Int'l Conf. on Research and Development in Information Retrieval*, pp. 256–263, 2000.
- [7] T. Furukawa and M. Kuzunishi, "Hierarchical Classification of Heterogeneous Data," *Proc. IASTED Int'l Conf. on Databases and Applications (DBA2005)*, pp. 252–257, 2005.
- [8] N. Ghamrawi and A. McMallum, "Collective Multi-Label Classification," *Proc. Int'l Conf. on Information and Knowledge Management (CIKM'05)*, pp. 195–200, 2005.
- [9] D. Koller and M. Sahami, "Hierarchically Classifying Documents Using Very Few Words," *Proc. Int'l Conf. on Machine Learning*, pp. 170–178, 1997.
- [10] M. Kuzunishi and T. Furukawa, "Representation for Multiple Classified Data," *Proc. IASTED Int'l Conf. on Databases and Applications (DBA2006)*, pp. 135–142, 2006.
- [11] A. Sun and E. Lim, "Hierarchical Text Classification and Evaluation," *Proc. IEEE Int'l Conf. on Data Mining (ICDM2001)*, pp. 521–528, 2001.
- [12] K. Toutanova, F. Chen, K. Popat, and T. Hofmann, "Text Classification in a Hierarchical Mixture Model for Small Training Sets," *Proc. Int'l Conf. on Information and Knowledge Management (CIKM'01)*, pp. 105–112, 2001.
- [13] K. Wang, S. Zhou, and Y. He, "Hierarchical Classification of Real Life Documents," *Proc. SIAM Int'l Conf. on Data Mining*, pp. 1–16, 2001.
- [14] K. Wang, S. Zhou, and S. C. Liew, "Building Hierarchical Classifiers Using Class Proximity," *Proc. Int'l Conf. on Very Large Data Bases (VLDB'99)*, pp. 363–374, 1999.