

# 博 士 学 位 論 文

内容の要旨および審査結果の要旨

第 3 3 号

2 0 2 6 年

東 京 国 際 大 学

## は し が き

本号は、学位規則（昭和 28 年 4 月 1 日文部省令第 9 号）第 8 条による公表を目的とし、2026 年 8 月 31 日に本学において博士の学位を授与した者の論文内容の要旨および論文審査結果の要旨を収録したものである。

学位記番号に付した甲は、学位規則第 4 条第 1 項（いわゆる課程博士）によるものであり、乙は同条第 2 項（いわゆる論文博士）によるものである。

目 次

| 学位記番号   | 学位の種類                 | 氏 名                | 論 文 題 目                                                                                                       | 頁 |
|---------|-----------------------|--------------------|---------------------------------------------------------------------------------------------------------------|---|
| 甲第 65 号 | 博 士<br>(デジタル<br>経営革新) | WU<br>QIANTIANHONG | Clustering Sparse Product-Interaction Data: A<br>Cosine- vs. Hamming-Distance Approach to<br>Churn Prediction | 1 |

|             |                                                                                                         |
|-------------|---------------------------------------------------------------------------------------------------------|
| 氏名・（本籍地）    | WU QIANTIANHONG（中国）                                                                                     |
| 学 位 記 番 号   | 甲第 65 号                                                                                                 |
| 学 位 の 種 類   | 博 士 （デジタル経営革新）                                                                                          |
| 学位授与の日付     | 2026 年 8 月 31 日                                                                                         |
| 学位授与の要件     | 学位規則第 4 条第 1 項該当                                                                                        |
| 学 位 論 文 題 目 | Clustering Sparse Product-Interaction Data: A Cosine- vs. Hamming-Distance Approach to Churn Prediction |
| 論文審査委員      | （主査） 教授 Nora Sharkasi                                                                                   |
| 論文審査委員      | （副査） 教授 Jay Rajasekera                                                                                  |
| 論文審査委員      | （副査） 教授 Parag Kulkarni                                                                                  |

## I. 論文内容の要旨

### DOCTORAL DISSERTATION ABSTRACT

Marketing datasets are often highly sparse, particularly in e-commerce environments where customers interact with only a limited subset of product categories. Although Product Hierarchy Modeling PHM provides a meaningful structure for representing category-level engagement, limited research has examined whether root-level category interaction patterns contain predictive signals for customer churn. Conventional clustering approaches addressing sparse interaction data often rely on binarized categorical representations, such as Hamming-distance K-modes clustering, or on alternative frameworks including hurdle-style clustering, embedding-based clustering, and Cosine-based hierarchical clustering, each involving trade-offs between computational tractability, interpretability, and the preservation of behavioral engagement intensity. This study examines how similarity representation influences segmentation and churn prediction in sparse behavioral data. Specifically, Cosine-based similarity is introduced as an intensity-preserving alternative to the mismatch-based Hamming distance traditionally used in clustering sparse categorical interactions. The resulting clusters are incorporated into machine learning churn prediction models alongside traditional Recency–Frequency–Monetary RFM indicators. Empirical analyses using B2B and B2C e-commerce datasets show that Cosine-based clustering produces stronger cluster cohesion, clearer separation, and more interpretable engagement patterns than Hamming-based clustering. In churn prediction tasks, models incorporating Cosine-derived PHM segments achieve higher recall and fewer false negatives, indicating improved identification of customers at risk of churn.

Keywords: Cosine distance, Hamming distance, clustering, Machine Learning Classification, Gradient Boosting, Logistic Regression, Easyensemble, Product Hierarchical Modeling PHM, Recency Frequency Monetary RFM.

## II. 論文審査結果の要旨

審査対象者： 学籍番号 23177901      氏名 WU QIANTIANHONG  
論文題目： Clustering Sparse Product-Interaction Data: A Cosine- vs. Hamming-Distance Approach to Churn Prediction

論文審査委員会

審査委員長    Nora Sharkasi  
審査員            Jay Rajasekera  
審査員            Parag Kulkarni

### Summary of Dissertation Review Result

The primary objective of this thesis is to address a critical methodological problem in marketing analytics: how to appropriately analyze highly sparse and zero-inflated customer interaction data for clustering and customer segmentation. Specifically, the study investigates: (1) how the choice of distance measure affects clustering performance when product-category interaction data contain a large proportion of zero values; (2) whether Cosine-distance-based clustering can preserve behavioral intensity and provide more meaningful customer segments than conventional Hamming-distance-based approaches; and (3) whether incorporating Product Hierarchy Modeling (PHM)-based customer clusters into the conventional Recency–Frequency–Monetary (RFM) framework can enhance the prediction of customer churn and improve the predictive performance of machine-learning models. The thesis directly addresses the problem that sparse product-interaction data can distort similarity measures and reduce model stability, while conventional binarization approaches may discard important information concerning the intensity of customer engagement.

A major contribution of the research is therefore its treatment of distance measurement under data sparsity. Rather than reducing customer-category interactions to simple

presence-or-absence indicators, the study applies Cosine distance to retain the proportional structure of customer engagement and systematically compares it with the conventional Hamming-distance K-modes approach. The proposed Product Hierarchy Modeling framework clusters customers according to their interactions with root-level product categories, thereby creating an interpretable behavioral segmentation layer that complements the traditional RFM dimensions.

The empirical findings demonstrate the importance of this approach for both segmentation and churn analytics. Cosine-distance clustering produced substantially stronger cluster cohesion and separation than the Hamming-based approach, while the resulting PHM clusters contributed additional behavioral information when incorporated into Logistic Regression, Gradient Boosting, and EasyEnsemble churn-prediction models. In particular, the Cosine-based configurations generated higher true-positive rates and fewer false negatives, demonstrating that category-hierarchy-based behavioral segmentation can improve the ability of the conventional RFM framework to identify customers at risk of churn.

Importantly, this research has also been developed into a peer-reviewed manuscript for the Journal of Marketing Analytics, reflecting the relevance of the work to contemporary research in data-driven marketing and customer analytics. The manuscript specifically positions the contribution as a sparsity-aware segmentation and prediction framework for high-dimensional product-category interaction data.

Wu, Q., Sharkasi, N. Clustering sparse product-interaction data: a cosine-versus hamming-distance approach to churn prediction. J Market Anal (2026). <https://doi.org/10.1057/s41270-026-00483-1>

博士学位論文 内容の要旨および審査結果の要旨（第 32 号）2026 年

---

2026 年 9 月 29 日

編集・発行 東京国際大学  
〒350-1197 埼玉県川越市的場北 1-13-1

---