

# Self-Clustering for Identification of Customer Purchase Behaviours

Guillem Lefait<sup>\*,\*\* ,\*\*\*</sup>, Gilles Goncalves<sup>\*\*</sup>, Tahar Kechadi<sup>\*</sup>

<sup>\*</sup>UCD, Belfield, School of Computer Science and Informatics, Dublin 4, Ireland

guillem.lefait@ucd.ie, tahar.kechadi@ucd.ie

<sup>\*\*</sup>Univ Lille Nord de France, F-59000 Lille, France

<sup>\*\*\*</sup>UArtois, LG2IA, F-62400, Béthune, France

gilles.goncalves@univ-artois.fr

**Abstract.** La segmentation d'une base client peut avoir différents objectifs et plusieurs segmentation peuvent être utiles pour décrire les clients ou pour s'adapter avec les stratégies commerciales d'une entreprise. Dans ce papier, nous présentons un schéma expérimental visant à proposer un ensemble de segmentations alternatives. Ces segmentations sont produites sur des données réelles par la transformation des données initiales, la génération et la sélection de différentes segmentations.

## 1 Introduction

Clustering consists in the creation of groups, such as objects inside same groups are very similar and objects of different groups are very dissimilar Xu and Wunsch (2005). Clustering is used for different objectives : to explore the data set, to condense it into a small set of representative points or to organise the data.

Segmentation is the ability to recognise groups of customers who share the same, or similar, needs McDonald (1996). Not all the customers in a broadly-defined market have the same needs, therefore the segmentation enables companies to provide specific products or services to different segments.

The use of clustering to automatically provide a segmentation is not recent and has been performed for two main objectives : 1) to identify groups of entities that share certain common characteristics and 2) to better understand buyer behaviours by identifying homogeneous groups of buyers Punj and Stewart (1983). However, there are different challenges to address when using clustering to perform the segmentation : which data to select, how many clusters to produce and how to evaluate the clustering results.

Very few solutions have been proposed to evaluate the quality of the customer segmentation. Manual investigation is often the solution used to assess the relevance of the clusters Aggelis and Christodoulakis (2005). In Cheng and Chen (2009), the segmentation result is assessed by the accuracy to predict loyalty of unknown customers. In Chang et al. (2009), sales forecasting is used on the segmentation result.

In this paper, we are interested to create, select and evaluate clustering results that will be presented to experts. The idea presented in this paper is based on the systemic search