User-Centric Similarity Search

ABSTRACT:

User preferences play a significant role in market analysis. In the database literature there has been extensive work on query primitives, such as the well known top-k query that can be used for the ranking of products based on the preferences customers have expressed. Still, the fundamental operation that evaluates the similarity between products is typically done ignoring these preferences. Instead products are depicted in a feature space based on their attributes and similarity is computed via traditional distance metrics on that space. In this work we utilize the rankings of the products based on the opinions of their customers in order to map the products in a user-centric space where similarity calculations are performed. We identify important properties of this mapping that result in upper and lower similarity bounds, which in turn permit us to utilize conventional multidimensional indexes on the original product space in order to perform these user-centric similarity computations.

INTRODUCTION:

ESTIMATION of the similarity between objects is a fundamental operation in data management. For instance it is used to find pages or documents with similar words over the web or in order to detect customers with abnormal behavior based on the products they buy. Moreover, similarity computations can be performed for the detection of similar conversations and comments between users of the social networks (i.e. comments on Facebook, tweets on Twitter).

Many different similarity metrics have been proposed for evaluating the similarity between two data items, such as the Euclidean distance and the cosine similarity. Such metrics suggest that the similarity between data items is computed based on their attributes, without taking into consideration users' opinions. For example, in business analysis the products are represented as points, defined by their attributes values. The closer two products are to each other according to the selected metric, the more similar they are.