Facing the Product Reclassification Challenge

Sven Abels, Axel Hahn

University of Oldenburg, Business Information Systems
Ammerländer Heerstr. 114-118, 26111 Oldenburg, Germany
{abels|hahn}@wi-ol.de

Abstract: Classification systems are often used in the e-Business domain to arrange products into different categories (classes). Within this field, the reclassification deals with the challenges that occur whenever two or more product catalogs are based on different classification systems. In this situation, it is necessary to re-arrange the product into another class-structure in order to enable a semantic exchange of data. This paper introduces the reclassification challenge and it outlines an idea for performing a semi-automatic reclassification of products.

1 Motivation and Problem Description

Classification systems are used to arrange products into categories (classes). Those categories are independent from the product catalog in which the products are stored (compare to [LSD02]). They therefore offer an independent hierarchical structure of classes. Well known examples are, e.g., eCl@ss or UNSPSC (see [Ec05], [Un01]). Using classification systems allows adding a special class-code to products in order to indicate their position in the classification system. For example, the code “41-01-10-03” might mean that a product is contained in the class “toy” of a classification system. Advantages when adding classification information to product data are, for example, the easy localization of similar products. The use and exchange of catalog data, in which products are classified based on a certain classification system, is usually seen as a standardization problem (cf. [He03]). As long as one common standard such as eCl@ss 4.0 or UNSPSC 5.0 is used by all participating parties, an exchange and interpretation of classification systems is easily possible and meaningful.

This text deals with the problematic situation in which it was not possible to agree on a common standard. In this paper, we assume that we have to deal with two software applications, exchanging product data and we assume that those systems are based on different classification systems (see [Sc01] for a similar problem description). For an example, a product catalog, produced by a system might be classified using eCl@ss 4.0 but another system (i.e., an e-procurement solution), might expect UNSPSC 5.0 based data, which was not supported by the first system that created the catalog. In this case, a so called reclassification of product data is necessary (compare to [BM04]). This process means that each product has to be transferred from its old standard (eCl@ss 4.0) into the new one (UNSPSC 5.0). Basically, such a reclassification is necessary, whenever (a) different classification standards are used (eCl@ss <-> UNSPSC), (b) different versions
of the same classification system are used (UNSPSC 5.0 <> UNSPSC 7.0) or (c) whenever products have to be classified and updated based on more than one standard simultaneously (in order to simplify and to speed-up this update process). -- Even if a product has not been classified using a standardized classification system, it will in many cases be located in an ordered structure within the catalog because in most catalogs, product data is located within product groups. Those product groups can also be interpreted as some kinds of classification systems [Na03], which means that even in these cases, a reclassification and not just a classification has to be performed.

2 Related Work

Existing classification solutions may of course be reused within the reclassification area since the reclassification is nothing else but a special kind of classification with additional information from the old (existing) class structure. Examples, that could be applied for product classification are, for example, Naïve Bayes or SVM based approaches (see [SKT04], [Wi00], [DC00]). Those machine learning approaches have, however, the disadvantage that trainings data is necessary. An alternative approach is the usage of keyword-based analyzers that compare keywords, extracted from the product descriptions, to the class descriptions of the classification system. Furthermore, special product-related approaches could be applied such as, e.g., GoldenBullet [Di02], which was developed for classifying product data into classification systems.

An alternative way is the usage of a mapping between both classification systems. This means that a mapping has to be produced for all classes in order to define that, e.g., class XY of system 1 is identical to class AB in system 2. In [BGV02], Bergamaschi, Guerra and Vincini introduce a way of determining such a mapping automatically. This can be interpreted as some kind of integration between two different ontologies. A major problem of this approach is that it cannot be applied, wherever two classes cannot be mapped in a 1:1 or at least n:1 way. An 1:m situation means that a detailed analysis of products is necessary. For example, class XY in system 1 might be called “paper”. Classification system 2 might, however, need an additional breakdown into “white paper”, recycled paper”, etc. This can only be performed by analyzing the product data itself. The following section will introduce an approach for those cases.

3 Reclassification of Product Data

Approaches, developed for pure classification of product data, have in many cases an essential disadvantage within a reclassification scenario. They do in many cases only observe the description of the product, which means that the old (existing) classification information is not used for the classification process. Furthermore, most approaches do not interpret additional metadata extracted from the product catalog or extracted from old classification information such as, e.g., the price or the manufacturer information. The authors of this contribution argue that this data can provide significant additional information, which might be used to improve the overall process.
The interpretation of existing classification information enables an interpretation of additional information. This information can be extracted from (i) the class name and synonyms of the class name, (ii) the description of the class, (iii) product information from classification attribute lists, (iv) structural information from the class hierarchy (parent classes), (v) relations to other classes (i.e. neighbor classes), (vi) versioning information (e.g. migration from eCl@ss 5.0 to eCl@ss 5.1) and (vii) information from the new class of products that where formerly located in the same class as the analyzed product.

It is purpose of the authors to determine, how helpful an interpretation of existing classification information can be when reclassifying product data. Popular product classification systems such eCl@ss or UNSPSC contain several thousand classes. It is purpose to use this information to reduce the number of potential classes for a product as much as possible and afterwards to perform an interpretation of the other product information in order to determine the most probable class for a product. Hence, we suggest several steps for performing a reclassification of products. Each step will end with a list of possible classes by reducing the original list of potential classes. It is proposed to divide the reclassification process into the following three steps:

1. Interpretation of existing classification information
2. Interpretation of classification attribute lists and additional product meta data
3. Analysis of the products description with classification approaches

In the first step, existing classification information is interpreted. Together with the next step, this one is the main difference between a pure classification and a reclassification process. It considers (a) information from the arrangement of a product into the catalogs product group structure and (b) its location in other classification systems, which differ from the destination system. For example, if a product shall be classified based on UNSPSC 5.0 and is presently located in the product group “Soccer Equipment” and contains the assignment to the eCl@ss 4.1-class “Ball (promotion)” then that information is used as a basis for determining potential UNSPSC classes. This is performed by finding classes in the destination system that are very similar to this information (e.g. by looking for similar keywords, etc.).

The first step results in a list of classes that are similar to the existing class information. There are, however, problems when dealing with mismatched and overlapping classes under two different taxonomy systems. In order to solve this, the second and third steps are performed. In the second step, additional data is analyzed which can be extracted by the attribute list of the old classification information. Furthermore, additional metadata from the product data in the product catalog is interpreted such as, e.g., the price of a product. If there are, for example, two possible classes available for a product called “train” and “toy train” then the price information (e.g. 12.75 €) might be enough to perform a decision, if it is known that other products in the class “train” cost >10.000,- € per product. Additional metadata are, for example, the dimensions of a product of its manufacturer. For interpreting this data, a product should be realized as some kind of grid, in which the each product is described by several attributes, a product group, a set
of existing classification information and a link to other products located in the same product group. The following screenshot of the Apricot-system shows a representation of this. Within this figure, dark fields are representing the classification information, interpreted in the first step.

![Apricot System Screenshot](image)

In the third step, the products' descriptions are analyzed. In order to perform this analysis, the textual data is cleaned (normalized) at first by stemming all words and removing unnecessary stopwords (see [CR04]). Afterwards, the descriptions are compared to the potential classes, detected in the first two steps. The most probable classes are then added to a resulting list. This comparison could be performed with techniques such as the keyword analysis or machine learning approaches such as Naïve Bayes or SVM (as described in section 2).

### 4 Realization of the Approach and Future Activities

The Apricot-project [AH05] of the University of Oldenburg is currently under development in order to realize the approach, sketched in the last section. It will provide a semi-automatic way of reclassifying products located in product catalogs. Apricot supports the import of product catalogs in the popular BMEcat format (see [SKP01]) and
it then performs the three steps, described above. Afterwards, it asks the user to pick one of the suggested classes. In the current implementation, the first two steps are already finalized, which makes it possible to perform a first test and impression of their effectiveness based on a small amount of selected product catalogs. In those tests, a list of potential classes for a product was detected exclusively by looking at their old classification information as well as by looking at the product price. In about 76% of all cases, the correct class was included within the first 10 matches.

Further work will be performed by completing the implementation of the approach. This will include the realization of the third step (interpretation of product descriptions). It is also proposed to interpret more metadata such as, e.g., the dimension of a product if this data is available. An interesting idea might also be to consider the context when reclassifying products. It, for example, the information is given that a catalog deals with toys, then a train will most likely be a toy-train and not a big vehicle. – After finishing the implementation, an exhaustive evaluation of the approach is planned, which will be based on real-world product catalogs. First evaluation results of these steps are expected between November and December 2005.

References

[Ec05] eCl@ss: eCl@ss. The leading classification system. White Paper, http://www.eclass.de.