Interoperability of catalog based systems

Sven Abels, Axel Hahn

University of Oldenburg, 26111 Oldenburg, Germany, {abels | hahn}@ wi-ol.de

Abstract. This short paper gives an introduction into electronic product catalogs and classification systems such as eCl@ss or UN/SPSC used in this domain. The role of classification systems to foster interoperability of catalog based enterprise systems is explained as well as problems that occur. Afterwards an approach is presented that aims in compensating these problems with the help of a mediating system used to re-classify products.

1 Introduction

In recent years, the area of e-commerce has become an important domain for enterprises. Today, most enterprises are directly or indirectly involved in this domain, either by selling products online or by using e-procurement systems to effectively manage their purchases. Because of an increasing importance of e-commerce, a large number of different competing products have been developed as well as different data formats to manage e-commerce data.

Electronic product catalogs (EPCs) are used to exchange product information between different enterprises in the business-to-business (B2B) field. Modern electronic product catalogs contain at least (i) a number of different products, (ii) details for each product such as price information or the manufacturer and (iii) a number product groups into which the products are arranged. A detailed description and a broad literature survey is given by in [1]. Popular examples are BMEcat[2] or xCBL[5], which are based on XML.

2 Interoperability of catalog based systems

Using standard catalog formats, such as BMEcat, enables an easy collaboration between enterprises when exchanging product data. There are many standard applications that are able to import and interpret catalog data, stored in these formats.

Many enterprises have to integrate more then one catalog into the own system. For example, in e-procurement systems of enterprises, designed to support the electronic procurement of goods, products from a large number of suppliers are integrated into one system (see [3], [4]). Other companies might have to offer products from multiple suppliers in an own web-shop.

There are several serious problems when integrating catalogs from more then one supplier into the own system. Basically, there are two major problems:

1. A different syntax and semantic of the data model (e.g. BMEcat vs. xCBL)
2. Different taxonomies and terminologies of the catalogs itself

The first problem is the usage of different catalog formats, which are incompatible. For example, it might be possible that a supplier offers his product in xCBL while the vendor's system expects it to be in the BMEcat format. An appropriate solution for this problem is to develop a simple converter that performs a conversation of the xCBL catalog into BMEcat. The second problem is, however, much more complicated then the first one and it is independent from the catalog format. The problem is that each product catalog might have its own product groups to arrange products. Basically, they differ in (i) their taxonomy, e.g. by having different subgroups for a category 'paper', (ii) their terminology, e.g. by using 'paper' and 'writing material' for an identical category and (iii) their language or spelling.

In order to solve those problems, classification systems were defined. A classification system is used „to assign each product to a product group corresponding to common attributes or application areas"[6]. Popular classification systems are eCl@ss [7] or UNSPSC [8]. Those systems offer a set of categories ("classes"), which are ordered hierarchically. A product can easily be assigned to a category by adding the category string to its product data. Classification systems can help to integrate products into existing catalogs and systems. There are, however, serious problems that prevent interoperability of different catalog based systems although a common catalog format was chosen and although classification systems were used. These problems are caused by the different standards in the domain of classification systems because there are several classification systems in this domain, which are incompatible. Hence, a reclassification of product data is necessary, whenever two e-commerce systems are using different classification systems, i.e. a conversation from eCl@ss information into UNSPSC values.

3 Reclassification of product data

Re-classifying product data as explained in the last section is not an easy task because a simple mapping between the categories of both classification systems is not possible in many cases (see e.g. [9], [10]). For example the first classification system might have a category called *Paper* in the main category *office materials.* The destination system might now need an additional break down into *White Paper*, *Recycled Paper*, etc. Hence, additional information is needed to re-classify all data correctly.

When looking at related problems, we can identify two related research areas:

1. Model transformation approaches, used to transform different models (c.f. [11]).
2. Typical classification approaches such as Bayes or a Vector-based classification [12].

Applying model transformation approaches for the reclassification is in most cases not enough because in these approaches, only the model itself is considered. In many cases, the models of different classification systems are almost identical or at least very similar but their contents such as, e.g., the name of the categories, differ completely.

Many typical ("traditional") classification approaches fail in the area of (re-)classifying product data since there are many different and similar classes (eCl@ss has over 24000 different classes). Existing solutions try to analyze products descriptions to extract keywords which are used to assign a product to a class. An example is given in [12] where Ding et al. indicate, to achieve a precision of 78% with one a Naïve-Bayes classification to classify 40% of the products, while they used the other 60% as training data for the algorithm. Existing solutions are designed for the classification of product data only. The authors argue that within the reclassification of product data, a number of additional information can be considered which can significantly improve existing approaches.

For example a product might be stored in an electronic catalog within a group "office material" with the name and description "Writingstar 4000+, 80g". Without looking at existing classification information, it will be hard to classify this product into e.g. the eCl@ss system. This task is much easier if existing classification information is interpreted. In the given example, there might be the UNSPSC code 14111511, which stands for 'writing paper'.

4 Conceptualization

The authors propose a mediator-like system to re-classify product data with respect to existing classification information. This system is supposed to modify a product catalog and re-classify all product data before the catalog is forwarded to the destination enterprise system (e.g. to an e-procurement system).

The reclassification process is to be performed in two steps.

1. The first step is performed by analyzing existing classification information and building a set of classes that could be chosen for the product in the new classification system. Since most classification systems are hierarchically ordered, it is in most cases easily possible to find such a set of matching classes. For example, if the existing classification information is called "writing paper" with the parent class "paper", then the new classification system will be searched for classes, which contain "writing paper". Those classes are added to a set of possible results. If no class was found, then all classes containing "paper" will be added.
2. The second step is to narrow down the set of possible classes by analyzing the product's description. This is performed in a similar way of existing classification solutions, which means that the product description is analyzed and keywords are extracted. Many approaches are using a machine learning approach to enrich their data with new keywords once, a product was classified correctly. A detailed description of such a classification is given in [13] or [12].

The suggested approach dramatically cuts down the number of categories that have to be analyzed in the analysis process. The following figure shows the suggested procedure graphically.

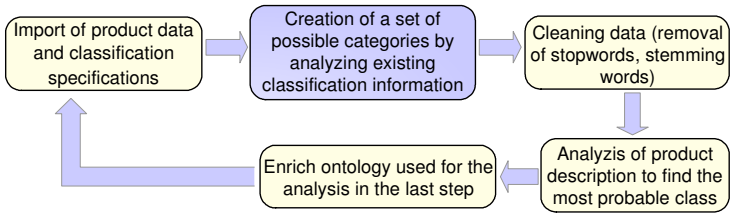

Figure 1: Considering existing classification information

The white ellipse show necessary steps for classifying products while the hatched ellipse shows the reclassification step, which filters the set of classes that can be chosen to classify the product. Without this step the analysis process would have to select a class from all possible classes, which is usually a high amount of categories, e.g. eCl@ss contains over 24.000. First tests have shown that the use of existing classification information can cut down this list to filter between 95% and 98% in good scenarios.

6 References

1. A. Patankar, A. Segev: An Extensible Catalog Management Framework, Fisher Center for Information Technology and Marketplace Transformation, Working Paper, 2002
3. T. Renner, V. Schmitz.: Specification BMEcat, V 1.2, www.bmecat.org, last access: 01/05, 2001
4. D. Neef: e-Procurement: From Strategy to Implementation. Financial Times Prentice Hall, 2001.
5. J. Hentrich: B2B-Katalog-Management (E-Procurement & Sales im Collaborative Business). Galileo Business, 2001.
6. xCBL: Structure reference, Version 4.0, http://www.xcbl.org/xcbl40, Last access: 12/04, 2003
7. J. Leukel, J., V. Schmitz, F.-D. Dorloff: Modeling and Exchange of Product Classification Systems using XML, in: Proceedings of the 4th IEEE International Workshop on Advanced Issues of E-Commerce and Web-based Information Systems, 2002
8. eCl@ss: eCl@ss White Paper, V0.6, 2001, http://www.eclass.de, last access: 17th Sept. 2004
9. UNSPSC: Why Coding and Classifying Products is Critical to Success in Electronic Commerce, Using the UNSPSC, White Paper, Granada Research. 2001
10. E. Schulten et al. : The E-Commerce Product Classification Challenge. IEEE Intelligent Systems Magazine, special issue on Intelligent E-business, (July/August 2001), 2001.
11. Y. Ding et al: The role of ontologies in e-Commerce. In: S.Stab & R. Studer (eds.) Handbook on Ontologies,Springer, 2003
12. N. Silva; J. Rocha: E-Business Interoperability through ontology semantic Mapping. E-business and Virtual Enterprises, 2003.
14. Y. Ding et al. : GoldenBullet: Automated Classification of Product Data in E-commerce. In: Proceedings of Business Information Systems 2002 (BIS2002), Ponznan, Poland, 2002.
15. S. Abels, A. Hahn: Conception of a framework for the combination of heterogeneous methods for re-classifying product data in the e-business domain. In: Proceedings of the 26th McMaster World Congress. Hamilton, Canada. 2005.