# Semantic and Syntactic Matching of e-Catalogues
## *Using Vector Space Model*

Ahmad Mehrbod, Aneesh Zutshi and António Grilo
*Unidade de Investigação em Engenharia Mecânica e Industrial (UNIDEMI), FCT/UNL, Campus de Caparica, Portugal*

Keywords: e-Catalogue, Semantic and Syntactic Integration, Vector Space Model.

Abstract: E-catalogues play a key role in e-procurement. Matching a product request from a buyer with products e-catalogues, helps companies to find business partners in e-marketplaces. Plethora of e-catalogue standards and enterprise specific formats is one of the major challenges in the matching process. In addition to this syntactic heterogeneity, the same product concept can be expressed using different words that causes semantic diversity in e-catalogues. In this paper, we propose a flexible approach to e-catalogue matching using Vector Space Model. Each e-catalogue is interpreted syntactically in its schema and semantically in an ontology that is developed based on its product classification system. Schemas and ontologies are added to the VSM by adding the syntax of the structure and semantic of the ontology to the indexing and searching mechanisms. The matching process uses the syntactic and semantic metadata for interpreting each e-catalogue as much as the information is available for the system, but is not dependent on this information.

## 1 INTRODUCTION

E-catalogues are electronic representations of information about the products and services of an enterprise (Huang and Huang, 2005). They play a critical role in e-procurement marketplaces. E-catalogues can be used in both the tendering (pre-award) and the purchasing (post-award) processes. Companies use e-catalogues to exchange product information with business partners. Suppliers use e-catalogues to describe goods or services that they offer for sale. Meanwhile buyers may use e-catalogues to specify the items that they want to buy (Ghimire et al., 2013a) (Ghimire et al., 2013b).

Matching a product request from a buyer with product e-catalogues that have been provided by the suppliers, helps companies to reduce the efforts needed to find partners in e-marketplaces(Lee et al., 2007).

This research is based on a research and development project in an e-procurement provider company and provides a framework for matching various e-catalogues originating from suppliers and buyers in the e-procurement platform. In an e-procurement e-marketplace at least two scenarios for finding similar e-catalogues are possible (Ghimire et al., 2013b) (Grilo et al., 2013a). First a buyer who makes a call for tender needs to select some suppliers based on their e-catalogues in order to send the invitation. The second scenario occurs when a supplier searches to find opportunities in e-marketplace. A supplier may upload a product e-catalogue to the search interface in order to find similar call for tenders. In the search scenario, a user has an e-catalogue and seeks similar e-catalogues in the platform.

## 2 RESEARCH PROBLEM

The large variety of e-catalogue formats which are used by various companies is a major challenge in the matching process. Since each business actor may use a different structure, classification and identification code for describing e-catalogues, it is not easy to match a product with the e-catalogue requested by another partner (Lee et al., 2007). This heterogeneity makes it difficult and time-consuming to integrate and query e-catalogues (Chen et al., 2010a).

While, there are too many different standards for e-catalogues and product classifications in use, often companies do not follow standard formats and prefer to have their individual structures (Chen et al., 2010b). Hence we often encounter a plethora of catalogue formats (Grilo and Jardim-Goncalves,


Mehrbod A., Zutshi A. and Grilo A..
Semantic and Syntactic Matching of e-Catalogues - Using Vector Space Model.
DOI: 10.5220/0005115302240229
In *Proceedings of the 11th International Conference on e-Business* (ICE-B-2014), pages 224-229
ISBN: 978-989-758-043-7

2013) ranging from unstructured text to well-structured XML documents. This diversity results in the syntactic heterogeneity of e-catalogues.

Syntactic diversity is only one side of the heterogeneity problem of matching e-catalogues. Other and yet more complicated side of this problem is the semantic diversity of e-catalogues. The same product concept can be expressed using different keywords, classifications or expressions. As result, different users may use different terms to express a same product concept, which makes the matching process get different result when facing a synonym query (Chen et al., 2010b).

# 3 STATE OF THE ART

The heterogeneity of e-catalogues which come from various sources (Grilo et al., 2013b) causes difficulty in the matching process. Generally we encounter with two aspects of heterogeneity in e-catalogues which are semantic and syntactic diversity. Syntactic heterogeneity is the result of different document structures and catalogue formats. Semantic heterogeneity refers to the different meanings of the words in various contents(Lee et al., 2007).

In order to encounter with the integration problem of e-catalogues, several approaches and methods have been proposed:

## 3.1 Standardization

In order to avoid the semantic diversity, classification systems such as CPV, UNSPSC and eCl@ss try to standardize the terms used for describing goods and services which are the subject of procurement. Using a common classification system for products and services enables reliable and efficient exchanges of product data across organizations. Additionally, e-catalogue standards such as UBL, BMEcat and cXML recommend using of these classification systems and furthermore propose common data structures for unifying e-catalogue schemas usually for exchanging purposes.

However, catalogue standards and classification systems are not sufficient to meet all the requirements of data exchange. Consequently, often enterprises do not follow standard formats and prefer to have their individual structures(Chen et al., 2010b). Also, variety of standards makes it impractical to reach the classification and schema unification goal. These standards differ in addressed markets, capabilities to represent product information, market acceptance, and standardization processes (Schmitz et al., 2005). This problem is more visible in multi-source e-marketplaces (Ghimire et al., 2013a) (Grilo and Jardim-Goncalves, 2013). There are at least 25 standards relating to e-catalogue and product classification systems, and thousands of enterprise products databases and e-commerce sites(Kim et al., 2007)(Chen et al., 2010b)(Schmitz et al., 2005).

## 3.2 Uniform Model

One traditional approach to solve the integration problem is to transform different formats into a uniform catalogue model (Ghimire et al., 2013b)(Ghimire et al., 2013a)(Chen et al., 2010a) that serves as reference format. In order to achieve this general model, these approaches formulate a formal model to represent various catalogues or select an existing standard as the general model. Then mapping functions have been designed that can handle the transformation of different formats into the uniform model.

But within this heterogeneous set of known or even unknown structures achieving a uniform structure for e-catalogues is usually not practical. Development of a uniform e-catalogue model requires precise and detailed understanding of each of the various formats of catalogues. However, there is always a chance to encounter with a new format which may cause difficulties in its interpretation. Furthermore for transformation to a uniform model, e-catalogues must be completely validated and in conformance to the expected format with no tolerance from format deviations.

## 3.3 Ontological Model

Syntactic interoperability alone does not handle all the problems of integration. Even if we get structured product information, it does not guarantee that we can interpret the content precisely when e-catalogues use different taxonomies. So, ensuring semantic interoperability is inevitable in the interpretation of product information(Kim et al., 2007). Therefore, several efforts such as (Huang and Huang, 2005), (Chen et al., 2010a) and (Lee et al., 2006) have encountered the problem by providing a universal ontological model for product data.

The purpose is to introduce generic attributes to design a universal ontology repository in order to facilitate e-catalogue sharing and interoperability (Chen et al., 2010b).The model is then used as a standard reference for e-catalogue transformation or development.

These ontologies are representation of products and services which include the definitions, properties, and relationships of the concepts that are fundamental to products and services(Lee et al., 2006). Usually these ontologies are constructed based on semantic concepts of either a product classification system or a product database. Many companies classify products according to generic or industry specific product classification standards, or by using proprietary category systems. Such classification systems often contain thousands of product classes that are updated over time. This implies a large quantity of useful product category information for e-commerce applications. Thus, instead of building up product ontologies from scratch, which is costly, tedious, error-prone, and high-maintenance, it is generally easier to derive them from existing product classifications (Stolz et al., 2014) (Lee et al., 2006).

Therefore these models have all the benefits of standard schemas and product classification systems and additionally improve accuracy of integration process using semantic relationships. But they also have the drawbacks of the standardization approach. It seems impossible to have a globally accepted reference model to create e-catalogues. Furthermore, transforming product specifications to such models not only is crucial and relies on manual efforts of domain experts, usually led to inadequate results.

### 3.4 Ontology Merging

Generally, this approach is similar to uniform model approach but targets the semantic integration instead of syntactic integration. In this approach instead of developing a universal ontological model to create e-catalogues or transform them to a common world, each e-catalogue is interpreted in its own ontological model. Developing various models for each type of e-catalogues is easier and more practical than creating a universal model for covering all kind of e-catalogues.

Due to different e-catalogue ontologies being generated from different data sources are heterogeneous, the key of semantic integration of e-catalogue in this way turns out to be the mapping and integration of catalogue ontologies (Chen et al., 2010a)(Chen et al., 2010b). Therefore different ontologies are combined into a new ontology that includes and reconciles all the information from the source ontologies according to semantic relations. For example (Kim et al., 2007) designed a product information mediation architecture by proposing ontology mapping algorithm regarding both taxonomy and attributes of underlying ontologies.

## 4 METHODOLOGY

### 4.1 Critique to Current Approaches

Based on two aspects of heterogeneity, syntactic integration and semantic integration of multi-source electronic catalogues have been attended to make e-catalogues interoperable. Both of them require integration of international product classification standards, enterprise product databases and product e-catalogue standards(Kim et al., 2007)(Chen et al., 2010b).

Some previous research, such as from (Ghimire et al., 2013a) and (Ghimire et al., 2013b) considered more syntactic integration, others such as (Kim et al., 2007) were more focused on semantic integration and some such as (Huang and Huang, 2005) studied both at the same time. But regardless of the semantic or syntactic dimension of the problem, the general solution in e-catalogue integration is to define a global model and convert e-catalogues to this uniform model. Therefore these traditional solutions either for semantic integration or syntactic integration are dependent on universal formal models.

### 4.2 Proposed Approach

Since in the area of e-catalogue we often encounter a plethora of formats, and developing a universal model is crucial, this research proposes to apply a more flexible approach to the e-catalogues matching problem. Vector Space Model (VSM) which is the base of many search techniques and document similarity methods can be applied to both semantic (Mukerjee et al., 2011) and syntactic (Manning et al., 2008) aspects of search problem. Although VSM has been used in various search problems, its application in matching e-catalogues is recent (Mehrbod et al. 2014). In the context of current research, VSM will be used to measure the syntactic and semantic similarity ratio of providers' e-catalogues with a buyer's e-catalogue.

Instead of developing semantic or syntactic models and combining them to universal models that try to cover all possible cases, the idea is to interpret each e-catalogue syntactically in its schema and semantically in an ontology that is made based on its product classification system. Schemas and ontologies will be added to the matching process by

adding the syntax of the structure and semantic of the ontology to the indexing and searching mechanisms of VSM. The matching process uses the syntactic and semantic metadata for interpreting each e-catalogue as much as the information is available for the system. Otherwise, it uses the basic mechanism of VSM for the unknown formats.

### 4.2.1 Vector Space Model

VSM uses occurrence of keywords or terms in documents to produce a table of vectors. Having a vector model of documents, mathematical vector operations can be applied to determine the similarity of a document with another one or with a search query. The simplest example is to use the deviation angle between vectors of frequent terms to calculate the relevance between text documents. While VSM is used to deal with flat textual data, it is being extended since the last two decades to treat complex structured and semi-structured data (Tekli et al., 2009).

### 4.2.2 Syntactic VSM

In order to encounter the syntactic heterogeneity using VSM, we considered three general cases for e-catalogues (Mehrbod et al. 2014). First, unstructured text such as PDF files which are common in online commerce. Second, structured or semi-structured documents which are unknown for the system such as enterprise specific formats. Third, structured standard documents which are known for the system such as cXML and UBL e-catalogues.

We used a Natural Language Processing tool to extract terms from e-catalogues. Then, a vector was made to represent occurrence of terms in each e-catalogue. E-catalogues that are similar to a given search query can be calculated by comparing deviation of angle between the vector of each e-catalogue and that of the query.

In order to associate the syntaxes in calculating similarity, levels of attributes in structured e-catalogues are also included in the term definition. Any structured or semi-structured document can be shown using a tree. Since content is distributed at different levels of the tree, location of a term in the tree is effective on the value of the term(Tekli et al., 2009) and should be considered in term extraction.

Generally, values of attributes are disregarded in term extraction from structured documents. This approach is useful for structure-only comparing of documents(Tekli et al., 2009). But in the context of product features, similarity measure is more sensitive to the values which have been saved in the e-catalogue structures. Therefore in matching process of e-catalogues, values are crucial and are even more important than structures. Consequently, we used a structure-and-content tokenization process (Manning et al., 2008) to define the terms.

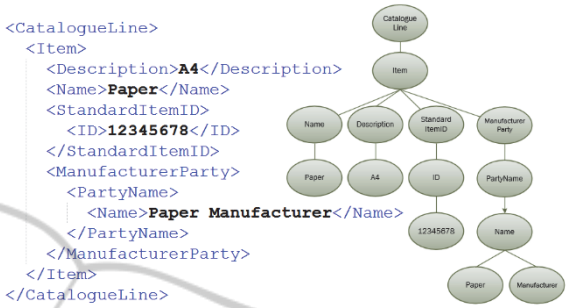


Figure 1: A part of a structured e-catalogue.

As an example, Figure 1 shows a portion of a UBL e-catalogue which is used by PEPPOL (Ghimire et al., 2013b). Table 1 shows the terms that are added to the vector of this e-catalogue for the value *paper* in the attribute *name*. Note that having values attached to all the terms helps search process to avoid matching documents with the same structure but different products.

In order to match e-catalogues that don't have completely the same structure with a lower similarity ratio, we divided the weight of a term by twice the number of nodes between the value and the attribute. For detail information please refer to (Mehrbod et al. 2014).

Table 1: Terms for *paper* in Figure 1.

| Value | Terms | Ratio |
|---|---|---|
| Paper | Paper | 1 |
| | Name/Paper | 1 |
| | Item/Name/Paper | 1 |
| | CatalogueLine/Item/Name/Paper | 1 |
| | Item/Paper | ½ |
| | CatalogueLine/Item/Paper | ½ |
| | CatalogueLine/Name/Paper | ½ |
| | CatalogueLine/Paper | ¼ |

Standard e-catalogues are source of diverse types of information. It usually includes general document data, product data and partners' data. This extra information can mislead product search process. Furthermore various attributes of product data can have different value in the matching process. For example classification code of a product has more value than a description in matching process. Hence we used tables of coefficients for known formats to adjust the impact of each attribute in matching process.

```
<CatalogueLine>
  <Item>
    <Description>0.5</Description>
    <Name>0.7</Name>
    <StandardItemIdentification>
      <ID>1</ID>
    </StandardItemIdentification>
    <ManufacturerParty>
      <PartyName>
        <Name>0</Name>
      </PartyName>
    </ManufacturerParty>
  </Item>
</CatalogueLine>
```

Figure 2: Coefficients for the sample e-catalogue.

Figure 2 shows the coefficients for the sample e-catalogue of Figure 1. These coefficients are values between 0 and 1 which are multiplied to the weights of terms. Undesired information such as partners' data can be simply excluded from matching process by putting 0 coefficients. Using this simple mechanism a new known structure can be easily added to the search system. Default values for all coefficients are 1 which reduces the status of an e-catalogue to an unknown structure for the matching process.

### 4.2.3 Semantic VSM

The general approach in semantic VSM is to expand terms using their synonyms in an ontology. As mentioned in the ontological modeling approach, many efforts have been done to build up product ontologies from existing classifications. These ontologies are rich sources of semantic information for interpreting product data. Recently, (Stolz et al., 2014) developed a generic, semi-automated method and tool for deriving OWL ontologies from product classification standards and proprietary category systems. Such approaches can be used to enrich product descriptions with information from existing data sources.

Hence it is straightforward to have an ontology for semantic presentation of each e-catalogue. In the case of standard formats, the relevant ontologies have been published by making ontologies from their product classification systems. For enterprise specific e-catalogues, ontologies can be provided by their companies using such available tools.

Then we have several overlapping ontologies in product data domain. Each one is a different abstraction and representation of the same or similar concepts. As mentioned many efforts have been done to integrate this ontologies using universal models that is crucial. In order to encounter the semantic heterogeneity, we want to empower our VSM approach (Mehrbod et al. 2014) using a simple, automatic and applicable ontology alignment process based on modeling ontologies in a vector space (Eidoon et al., 2008).

(Eidoon et al., 2008) developed a VSM for finding similar entities from one ontology to entities of another ontology without integrating them to a common model. In this method, vectors representing ontology concepts and properties are matched iteratively and their similarity degree is estimated.

By expanding e-catalogue matching process using this ontology alignment approach, we aim to find synonym or similar terms from one e-catalogue to ones of another e-catalogue. In this way the term-vector of each e-catalogue will be expanded by terms of all semantic concepts that exist in the e-catalogue. Each distinct concept and property represents one term in the vector space.

Adding these terms to the vectors and expanding similarity calculation with ontology alignment will enable the matching process to find semantically similar e-catalogues.



## 5 CONCLUSIONS

This research improves a G2B2B e-procurement platform by providing a semantic and syntactic matching mechanism for e-catalogues. Several companies use the e-procurement marketplace to purchase and sell products and services. In this process, e-catalogues can have an effective impact on finding the right business partners. Heterogeneity of e-catalogue formats that are used by various business players is a significant barrier to the application of the e-catalogues in searching procedures.

Several researches have been done to solve this multi-dimensional problem. In this paper we reviewed and categorized these approaches into four different classes and discussed their pros and cons. Most of these research works have some limitations for solving the problem including difficulties in implementation that make them difficult to implement.

In a previous work we had proposed an alternate solution to cope with this problem using flexible and practical method that is used by several search engines. We applied Vector Space Model to solve syntactic heterogeneity problems that we encounter in matching e-catalogues. In this paper we expanded the previous work to solve both syntactic and semantic aspects of the heterogeneity problem. First a brief review of the previous work on syntactic matching mechanism has been presented. The matching process had used a combinations of values

and attributes of structured documents to find the syntactic correlation of e- catalogues. A simple table of coefficients was proposed to specify the matching process for standard formats. This mechanism increases the search precision by removing unrelated information from the matching process and boosting weights of important attributes.

Then a practical approach to expand the matching mechanism was proposed. The proposed extension improves the search process with semantic matching using a simple, expandable and applicable ontology alignment approach. The aim is to use the benefits of all available ontologies and schemas but not to be dependent on them.

The e-catalogue matching approach is implemented as a prototype in the procurement platform. The results show the matching process is capable to match diverse formats of catalogues from various sources. The experimental results are not presented in this paper as the prototype is still on the final stages for full production deployment.



## ACKNOWLEDGEMENTS

The research of this work has been partially funded by project VortalSocialApps, co-financed by VORTAL and IAPMEI and the European Funds QREN COMPETE.

## REFERENCES

Chen, D., Li, X., Liang, Y., Zhang, J., 2010a. A semantic query approach to personalized e-Catalogs service system. J. Theor. Appl. Electron. Commer. Res. 5, 39–54.

Chen, D., Li, X., Zhang, J., 2010b. User-oriented intelligent service of e-catalog based on semantic web. 2010 2nd IEEE Int. Conf. Inf. Manag. Eng. 449–453.

Eidoon, Z., Yazdani, N., Oroumchian, F., 2008. Ontology matching using vector space. Adv. Inf. Retr. 4956, 472–481.

Ghimire, S., Jardim-Goncalves, R., Grilo, A., 2013a. Framework for catalogues matching in procurement e-marketplaces. In: Information Systems and Technologies (CISTI), 8th Iberian Conference on. pp. 1–6.

Ghimire, S., Jardim-Goncalves, R., Grilo, A., Beca, M., 2013b. Framework for inter-operative e-Procurement marketplace. In: Computer Supported Cooperative Work in Design (CSCWD), 2013 IEEE 17th International Conference on. pp. 459–464.

Grilo, A., Jardim-Goncalves, R., 2013. Cloud-Marketplaces: Distributed e-procurement for the AEC sector. Adv. Eng. Informatics 27, 160–172.

Grilo, A., Jardim-Goncalves, R., Ghimire, S., 2013a. E-Procurement in the Era of Cloud Computing. In: Proceedings of the 4th International Conference on Information Systems Management and Evaluation (Icime 2013). pp. 104–110.

Grilo, A., Jardim-Goncalves, R., Ghimire, S., 2013b. Cloud-Marketplace: New paradigm for e-marketplaces. In: Technology Management in the IT-Driven Services (PICMET), 2013 Proceedings of PICMET '13: pp. 555–561.

Huang, J.Z., Huang, G., 2005. Ontology-based e-catalog matching for integration of GDSN and EPCglobal network. IEEE Int. Conf. E-bus. Eng. 212–215.

Kim, W., Choi, D.W., Park, S., 2007. Agent based intelligent search framework for product information using ontology mapping. J. Intell. Inf. Syst. 30, 227–247.

Lee, J., Lee, T., Lee, S., Jeong, O., Lee, S., 2007. Massive Catalog Index based Search for e-Catalog Matching. 9th IEEE Int. Conf. E-Commerce Technol. 4th IEEE Int. Conf. Enterp. Comput. E-Commerce E-Services (CEC-EEE 2007) 341–348.

Lee, T., Lee, I., Lee, S., Lee, S., Kim, D., Chun, J., Lee, H., Shim, J., 2006. Building an operational product ontology system. Electron. Commer. Res. Appl. 5, 16–28.

Manning, C.D., Prabhakar, R., Schutze, H., 2008. Introduction to Information Retrieval. Cambridge University Press.

Mehrbod, A., Zutshi, A. & Grilo, A., 2014. A Vector Space Model Approach for Searching and Matching Product E-Catalogues. In J. Xu et al., eds. *Proceedings of the Eighth International Conference on Management Science and Engineering Management*. Advances in Intelligent Systems and Computing. Lisbon, Portugal: Springer Berlin Heidelberg.

Mukerjee, K., Porter, T., Gherman, S., 2011. Linear scale semantic mining algorithms in microsoft SQL server's semantic platform. Proc. 17th ACM SIGKDD Int. Conf. Knowl. Discov. data Min. - KDD '11 213.

Schmitz, V., Leukel, J., Dorloff, F., 2005. Do e-catalog standards support advanced processes in B2B e-commerce ? Findings from the CEN / ISSS workshop eCAT 00, 1–10.

Stolz, A., Rodriguez-Castro, B., Radinger, A., Hepp, M., 2014. PCS2OWL: A Generic Approach for Deriving Web Ontologies from Product Classification Systems. Semant. Web Trends Challenges SE - 43, Lecture Notes in Computer Science 8465, 644–658.

Tekli, J., Chbeir, R., Yetongnon, K., 2009. An overview on XML similarity: Background, current trends and future directions. Comput. Sci. Rev. 3, 151–173.