

Chapter 71

A Vector Space Model Approach for Searching and Matching Product E-Catalogues

Ahmad Mehrbod, Aneesh Zutshi and António Grilo

Abstract In e-procurement, companies use e-catalogues to exchange product information with business partners. The large variety of e-catalogue formats which are used by various companies make it difficult to match a product request from a buyer (buyer e-catalogue) with products e-catalogues. While, there are too many different standards for e-catalogues in use, often companies do not follow standard formats. Hence we often encounter a plethora of catalogue formats ranging from unstructured text to well-structured XML documents. One traditional approach to solve this problem is to convert different formats to a general common structure. But within this heterogeneous set of known or even unknown structures achieving a global structure is impractical. In this paper, vector space model has been used to measure the similarity ratio of providers' e-catalogues with a buyer's e-catalogue. Attributes of known structures and their values have been used as terms and their weights in the vectors to find the correlation of e-catalogues based on relationship of common tags. In order to associate the structures in calculating similarity, levels of attributes in xml documents are also included in the terms. Natural language processing is used to extract the same attributes from unstructured or unknown structured documents.

Keywords E-procurement · E-catalogue · Vector space model · Information retrieval

A. Mehrbod (✉) · A. Zutshi · A. Grilo
Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Campus de Caparica,
2829-516 Caparica, Portugal
e-mail: a.mehrbod@campus.fct.unl.pt

A. Mehrbod · A. Zutshi · A. Grilo
Unidade de Investigação em Engenharia Mecânica e Industrial (UNIDEMI), FCT/UNL,
Campus de Caparica, 2829-516 Caparica, Portugal

71.1 Introduction

E-catalogues play a critical role in e-procurement marketplaces. They 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 [5, 6].

Matching a product request from a buyer with products e-catalogs that have been provided by the suppliers, helps companies to reduce the efforts needed to find partners in e-marketplaces [12, 16].

The large variety of e-catalogue formats [21] which are used by various companies is one of the major challenges 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-catalog requested by another partner [16]. This heterogeneity makes it difficult and time-consuming to integrate and query e-catalogues [3].

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 [4]. Hence we often encounter a plethora of catalog formats ranging from unstructured text to well-structured XML documents.

One traditional approach to solve this problem is to transform different formats into a uniform catalogue model [3, 6, 12]. 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 [1]. 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.

This paper proposes to exploit Vector Space Model [18] which has been used by information retrieval systems to measure the similarity ratio of documents to match providers' e-catalogues with a buyer's e-catalogue. VSM uses occurrence of 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 it is used to deal with flat textual data (i.e. classical free text documents), IR is being extended, since the last two decades, so as to treat complex structured and semi-structured data [22].

The rest of this paper is structured as follows: Sect. 71.2 has a deeper view on heterogeneity of e-catalogues and reviews related works. In the Sect. 71.3 the proposed approach has been described in details. Experimental results are described in Sect. 71.4. We conclude with future work in Sect. 71.5.

71.2 Related Works

Regarding to the usage of e-catalogues in e-commerce, interoperability of e-catalogues (Catalogue integration) and personalization of e-catalogues are two main challenges which have been studied in the literatures. Although these challenges are related and many researchers studied both together, former is to match a search query with product e-catalogues and the latter is more focused on customizing e-catalogue selection based on user profile.

As mentioned, the heterogeneity of e-catalogues which come from various sources causes [8] 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 [16, 17].

In order to avoid this, diversity classification systems such as CPV¹, UNSPSC² and eCl@ss³ try to standardize the terms used for describing goods and services which are the subject of procurement. Additionally, e-catalogue standards such as PEPPOL⁴, BMEcat⁵ and ePRIOR⁶ recommend using of these classification systems and furthermore propose common data structures for unifying e-catalogue schemas usually for exchanging purposes.

But catalogue standards are not sufficient to the meet all requirements of data exchange [17]. Consequently, often enterprises do not follow standard formats and prefer to have their individual structures [4]. Not to mention variety of standards makes it impractical to reach the classification and structure unification goal. These standards differ in addressed markets, capabilities to represent product information, market acceptance, and standardization processes [21]. This problem is more visible in multi-source e-marketplaces [6–9]. There are at least 25 standards relating to e-catalogue and product classification, and thousands of enterprise products database and e-commerce sites [4, 14, 21].

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 the two require integration of international product classification standards, enterprise product database and product e-catalogue standards [4, 14]. Some researches such as [6] and [5] more considered syntactic integration and other such as [13, 14] and [11] more focused on semantic integration. But regardless of semantic or syntactic aspect of the problem, general solution in e-catalogue integration is to define a global model and convert e-catalogues to this uniform model or simply interpret them based on this reference model. For example [3] tries to formalize e-catalogues by offering an ontological model of e-catalogues. The main work in this kind of solutions is to introduce generic attributes to design e-catalogue ontology model [4]. Due to different e-catalogue ontology being generated from different data sources are heterogeneous, the key of semantic integration of e-catalogue turns out to be the mapping and integration of catalogue ontologies [4, 10].

Therefore these traditional solutions either for semantic integration or syntactic integration are dependent on universal formal models. As previously mentioned creating such general models has the following problems:

- Requires proper knowledge of the underlying catalogues' structures. But individual formats which are used by companies are usually unknown and usually there is chance to always encounter new formats. Lee et al. [16] provides a search index to match e-catalogues regardless of structures. But usually the structures contain valuable information. Our proposed approach exploits the structures as much as their details is known whilst is independent of structure. Structure independency provides the ability to match unstructured information as well as unknown structures. Information existing in structures is valuable and can be helpful in matching process.
- E-catalogues must be completely validated for conformance to their formats with no tolerance for format deviations. Since usually each structure is transformed to the general model, it has to be completely compatible with the structure which the convertor expects. Furthermore development of such convertors is crucial and time-consuming task.
- All the various matching cases in graphs or models must be predefined in matching algorithms. For example, Kwon et al. [15] tried to cover all the possible conditions in matching two structures. But within a heterogeneous set of structures always there is chance to encounter a new unconsidered condition.

71.3 Document Similarity in E-Catalogues

Since in the area of e-catalogue we often encounter a plethora of formats and developing an integrated model is crucial, this paper proposes to apply a more flexible model to e-catalogues search problem. Vector Space Model which is the base of many search techniques and document similarity methods can be applied to both semantic [19, 24] and syntactic [2, 18] aspects of search problem. Although semantic issues are also considered recently in document similarity techniques [19, 24], we will apply it to our approach in a future work. In this paper we will focus on overcoming syntactic diversity of e-catalogues including unstructured, unknown structured and known structured documents.

In an e-procurement e-marketplace at least two scenarios for finding similar documents are possible [5]. 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. Second scenario is when a supplier searches to find opportunities in e-marketplace. Supplier may upload a product e-catalogue to e-marketplace in order to find similar call for tenders.

Three general cases may be considered in syntactic interoperability of e-catalogues. 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 individual formats. Third structured standard documents which are known for the system such as PEPPOL e-catalogues. XML is one of the most common formats for exchanging structured and semi-structured data and also standard e-catalogues in B2B e-commerce [20]. Among 25 e-catalogue standards, 16 of them are based on XML [21]. Hence in the following sub-sections we will apply Vector Space Model to three groups of documents at the same time: unstructured text documents, XML documents and standard e-catalogues.

71.4 Unstructured Documents

In VSM, sets of keywords or terms have been extracted from documents and user queries. Then, a vector has been made to represent occurrence of terms in each document. If a term occurs in a document, its value or weight in the vector of the document is non-zero. Documents that are similar to a given query can be calculated by comparing deviation of angle between the vector of each document and that of the query.

Depending on the application, several methods have been proposed to define the weights. Keywords are commonly weighted in order to reflect their relative importance in the query or document at hand. The underlying idea is that terms that are of more importance in describing a given query or document are assigned a higher weight [18, 22]. For example one well-known method of weighting the terms is TF-IDF that takes into consideration both document and collection statistics [22].

Usually natural language processing techniques are utilized to extract important terms automatically from the documents and queries. Among various types of processing which can be applied to text documents, we used a Natural Language Processing tool to tokenize, lemmatize and remove stop words. Tokenization is to decide what constitutes a term and how to extract terms from raw text. Since our application is to match e-catalogues and term matching is more valuable than phrase matching, we filtered the stop words. Stop words are some of most common words such as the, is, at and so on. Lemmatisation is to convert the different inflected forms of a word to the lemma form so they can be analyzed as a single item [23].

71.5 Structured Documents

XML documents is widely used to represent structured information. Any structured or semi-structured document can be shown using XML files. Hence XML-based similarity becomes a central issue in the structured information retrieval. Since in conventional information retrieval, documents are unstructured data, Vector Space Model has been extended towards XML information retrieval [22]. Using these extensions we can represent structured and unstructured queries and documents in vector space model and compute matching between them.

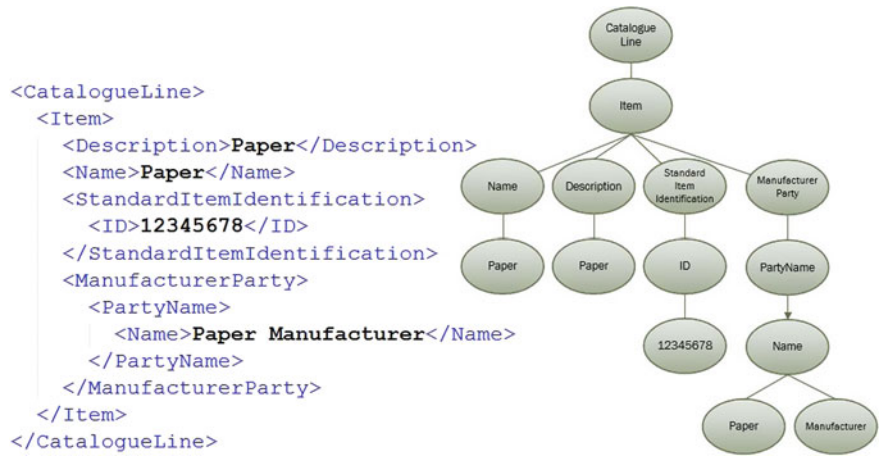


Fig. 71.1 A part of a standard e-catalogue (D1)

Hierarchical structure of XML documents are generally modelled as trees. In a traditional model, nodes of tree represent XML elements and are labelled with corresponding element tag names. 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 [22] and should be considered in term extraction.

In the comparison process generally values of attributes are disregarded. This approach is useful for structure-only comparing of XML documents [22]. 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 [18] to define the terms.

As an example, Fig. 71.1 shows a portion of a standard e-catalogue D1 which is used in PEPPOL [5]. In the matching process this e-catalogue should be similar to e-catalogue D2 in Fig. reffig:lurong04Ahamdfig2 that has the word paper. Moreover it should have a higher similarity ratio with document D3 in Fig. 71.2 that has the word paper in attribute name and even higher to document D4 in Fig. 71.2, that has paper in hierarchy of name and item and so on.

One way of doing this is to define a term as a value together with its position within the XML tree. Figure 71.2 illustrates this representation. We use all the sub-trees of a document that contain at least one value as terms [2, 18]. In other words, we first take each value (paper) as a term. This help the matching process to match this documents with unstructured documents or structured document such as D2 that have same values but in different structure. Next we add values with last level of their position (name/paper) to the terms. It helps matching process to increase similarity of D1 with structured documents such as D3. Then we continue adding levels of positions (item/name/paper) to the terms to root of tree. It keep increasing

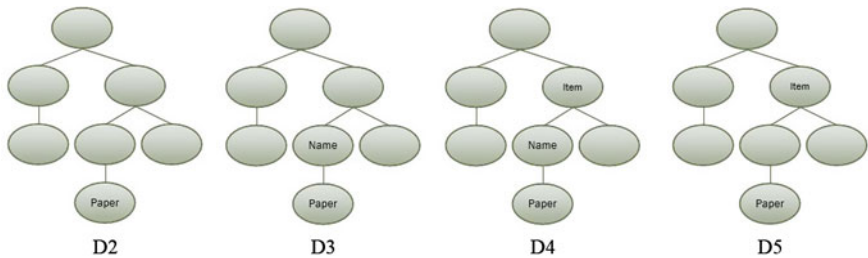


Fig. 71.2 Similar e-catalogues to D1

similarity of D1 with structures documents such as D4. Table 71.1 shows all possible terms for the tree of Fig. 71.1. Therefore D1 will have one common term with D2, two common term with D3 and tree common term with D4 which guarantees more ration of similarity for documents with resembling structures. Note that having value attached to all the terms helps search process to avoid matching documents with same structure but different products.

Documents such as D5 should have a lower matching ratio with D1 as compared to its matching ratio with D3 and D4, because the same value for an attribute (item) exists in both documents but not necessarily in same path order. Therefore the terms of Table 71.2 have also been added to the terms of D1 to cover this type of similarity. In order to decrease the similarity ratio for documents that match D1 using the terms of Table 71.2 instead of the terms of Table 71.1, we divide the weight of a term by twice the number of nodes between the value and the attribute. With this simple approach we don't have to change the similarity formula as proposed in [2] and [18].

71.6 Standard E-Catalogues

Standard e-catalogues are source of diverse types of information. For example a PEPPOL e-catalogue 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 a table of coefficients to adjust the impact of each attribute in similarity of known structures. These coefficients are values between 0 and 1 which are multiplied to the weight 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 structures can be easily added to the search system. Default value for all coefficients are 1 which reduces the status of an e-catalogue to an unknown structure for the matching process.

Table 71.1 All possible terms for D1

Value	Terms
Paper manufacturer	Manufacturer Name/manufacturer partyname/name/manufacturer Manufacturerparty/partyname/name/manufacturer Item/manufacturerparty/partyname/name/manufacturer Catalogueline/item/manufacturerparty/partyname/name/manufacturer Paper name/paper partyname/name/paper Manufacturerparty/partyname/name/paper Item/manufacturerparty/partyname/name/paper Catalogueline/item/manufacturerparty/partyname/name/paper
12345678	12345678 ID/12345678 Standarditemidentification/ID/12345678 Item/Standarditemidentification/ID/12345678 Catalogueline/item/Standarditemidentification/ID/12345678
Paper	Paper Description/paper Item/description/paper Catalogueline/item/description/paper
Paper	Paper Name/paper Item/name/paper Catalogueline/item/name/paper

Table 71.2 Additional terms for the last entry of Table 71.1

Value	Terms	Weight ratio
Paper	Item/paper	1/2
	Catalogueline/item/paper	1/2
	Catalogueline/name/paper	1/2
	Catalogueline/paper	1/4

71.7 Conclusion and Future Works

Based on many standards and data resource, query and search of e-catalogues is affected by integration problems [4]. Since IR-based methods are applicable to wide range of structured and unstructured documents which we encounter in matching e-catalogues, this paper proposes a vector space model approach to search e-catalogues. Furthermore these methods target loosely structured data, thus useful and generally exploited for fast simple structured search and retrieval [22].

Combinations of values and attributes of structured documents have been used to find the correlation of documents based on relationship of common tags. Then we have proposed a simple table of coefficients to specify the matching model for standard e-catalogues. This mechanism increases the search precision by removing

unrelated information from the matching process and boosting weights of important tags. In future these tables can be customized for users using a learning mechanism based on their profiles and search interests.

In order to test matching process we used a set of e-catalogues in various formats. First we utilized an open source full text search tool and a natural language analyser to extract terms from flat text files. Then we extended the search tool to consider the locational values of words in term extraction process when such information is available. Having tokenized all types of e-catalogues, we made term vectors for them and applied the coefficient tables to known structures.

With this tool, users will be able to search within the available files or simply upload their e-catalogues to find similar documents. This search mechanism allows users who prefer to specify the tag relations while searching [22] to get rid of using content-and-structure queries [19]. The experimental results show the matching process is capable to match diverse formats of catalogues from various sources.

Acknowledgments The research of this work has been partially funded by project VortalSocialApps, co-financed by VORTAL and IAPMEI and the European Funds QREN COMPETE, and also would like to thank Fundação da Ciência e Tecnologia for supporting the research center UNIDEMI through the grant PEST-OE/EME/UI0667/2011.

References

1. Benatallah B, Hacid MS et al (2006) Towards semantic-driven, flexible and scalable framework for peering and querying e-catalog communities. *Info Syst* 31(4–5):266–294
2. Carmel D, Maarek Y et al (2002) An extension of the vector space model for querying xml documents via xml fragments. *SIGIR Forum*
3. Chen D, Li X et al (2010) A semantic query approach to personalized e-catalogs service system. *J Theor Appl Electron Commer Res* 5(3):39–54
4. Chen D, Li X, Zhang J (2010b) User-oriented intelligent service of e-catalog based on semantic web. In: 2nd IEEE international conference on information management and engineering (ICIME), pp 449–453
5. Ghimire S, Jardim-Goncalves R et al (2013a) Framework for inter-operative e-procurement marketplace. In: 17th IEEE international conference on computer supported cooperative work in design (CSCWD 2013), pp 459–464
6. Ghimire S, Jardim-Goncalves R, Grilo A (2013b) Framework for catalogues matching in procurement e-marketplaces. *Iber Conf Inf Syst Technol Cist*.
7. Grilo A, Jardim-Goncalves R (2013) Cloud-marketplaces: distributed e-procurement for the aec sector. *Adv Eng Info* 27(2):160–172
8. Grilo A, Jardim-Goncalves R, Ghimire S (2013a) Cloud-marketplace: new paradigm for e-marketplaces. In: Technology management in the IT-driven services (PICMET), 2013 proceedings of PICMET, vol 13, pp 555–561
9. Grilo A, Jardim-Goncalves R, Ghimire S (2013b) E-procurement in the era of cloud computing. In: 4th international conference on IS management and evaluation (ICIME 2013), pp 104–110
10. Guo J (2009) Collaborative conceptualisation: towards a conceptual foundation of interoperable electronic product catalogue system design. *Enterp Info Syst* 3(1):59–94
11. Huang JZ, Huang G (2005) Ontology-based e-catalog matching for integration of gdsn and epccglobal network. In: IEEE international conference on e-business engineering, pp 212–215

12. Kim D, Kim J, Lee S (2002a) Catalog integration for electronic commerce through category-hierarchy merging technique. In: Proceedings of twelfth int work res issues data eng eng E-commerce/e-bus Syst RIDE-2EC 2002, pp 28–33
13. Kim D, Kim J, Lee S (2002b) Catalog integration for electronic commerce through category-hierarchy merging technique. In: Proceedings twelfth int work res issues data eng eng e-commerce/e-bus Syst RIDE-2EC 2002, pp 28–33
14. Kim W, Choi DW, Park S (2007) Agent based intelligent search framework for product information using ontology mapping. *J Intell Info Syst* 30(3):227–247
15. Kwon IH, Kim CO, KPK et al (2008) Recommendation of e-commerce sites by matching category-based buyer query and product e-catalogs. *Comput Ind* 59(4):380–394
16. Lee J, Lee T, et al. (2007) Massive catalog index based search for e-catalog matching. In: 9th IEEE international conference on e-commerce technology 4th IEEE international conference on enterprise computing e-commerce e-services (CEC-EEE 2007), pp 341–348
17. Leukel J, Schmitz V, Dorloff F (2002) Exchange of catalog data in B2B relationships. Analysis and improvement 2002(ICWI), pp 403–410
18. Manning CD, Prabhakar R, Schutze H (2008) Introduction to information retrieval. Cambridge University Press, Cambridge
19. Mukerjee K, Porter T, Gherman S (2011) Linear scale semantic mining algorithms in microsoft sql server's semantic platform. In: Proceedings of 17th ACM SIGKDD international conference on knowledge discovery and data mining—KDD'11, p 213
20. Schmitz V, Leukel J, Dorloff F (2003) Xml data modeling concepts in B2B catalog. In: Proceedings of IADIS international conference e-Society 2003 (ES 2003) 2003(Es), pp 227–234
21. Schmitz V, Leukel J, Dorloff F (2005) Do e-catalog standards support advanced processes in B2B e-commerce? The CEN/ISSS Workshop ECAT 00(C), pp 1–10
22. Tekli J, Chbeir R, Yetongnon K (2009) An overview on XML similarity: background, current trends and future directions. *Comput Sci Rev* 3(3):151–173
23. Turney P, Pantel P (2010) From frequency to meaning: vector space models of semantics. *J Artif Intell Res* 37:141–188
24. Widdows D (2008) Semantic vector products: some initial investigations. In: Second AAAI symposium on quantum interaction, March

Proceedings of the Eighth International Conference on
Management Science and Engineering Management
Focused on Computing and Engineering Management
Xu, J.; Cruz-Machado, V.A.; Lev, B.; Nickel, S. (Eds.)
2014, XXII, 798 p. 171 illus., 54 illus. in color., Hardcover
ISBN: 978-3-642-55121-5