The influence of knowledge in the design of a recommender system to facilitate industrial symbiosis markets

Guido van Capelleveen[a], Chintan Amrit[a], Devrim Murat Yazan[a], Henk Zijm[a]

[a]*Department of Industrial Engineering and Business Information Systems, University of Twente, The Netherlands*

Abstract

Industrial symbiosis aims to stimulate or enhance cooperation between industrial firms to utilize industrial waste streams from other industries and to share related knowledge, in order to achieve sustainable production. Recommenders can support industries through the identification of item opportunities in waste marketplaces, enhancing activities that may lead to the development of an active waste exchange network. To build effective recommendation, we study the role of knowledge in the design of a recommender that suggests waste materials to be used in process industries. This paper compares the performance of a knowledge based input-output recommender with a recommender based on association rules. The two recommenders are evaluated with real-world data collected through deploying surveys in a workshop setting. Our research shows that many data challenges arise when creating recommendations from explicit knowledge and suggests that techniques based on the concept of implicit knowledge may be preferable in the design of an industrial symbiosis recommender.

Keywords: Industrial Symbiosis, Recommender systems, Decision Support Systems, Input-output matching, Association-rule mining

*Corresponding author at: University of Twente, PO box 217, 7500 AE Enschede, The Netherlands, E-mail address: g.c.vancapelleveen@utwente.nl (Guido van Capelleveen)

1. Introduction

The reduction of waste emissions and primary resource use in resource-intensive industries is suggested as one of the critical pathways to accelerate sustainable development (European Environmental Agency, 2016). The European Union highlights Industrial symbiosis (IS) as a methodology that stimulates industries to become more sustainable. This policy follows from the vast amount of positive environmental effects generally associated with the methodology and its ability to scale at a European level (Laybourn and Lombardi, 2012). IS entails the identification and utilization of an organization's traditional secondary production process output that is generally considered as waste. Such waste can be used to substitute (part of) a primary resource (possibly after some pre-processing) in the production process of another organization, usually located in a different industrial sector (Chertow, 2000). Both researchers and practitioners recognize the impact of IS as a useful business opportunity and tool for eco-innovation. It involves "engaging diverse organizations in a network to foster eco-innovation and long-term culture change, creating and sharing knowledge through the network yielding mutually profitable transactions for novel sourcing of required inputs, value-added destinations for non-product outputs, and improved business and technical processes" (Lombardi and Laybourn (2012), page 28). New IS opportunities are mostly identified using facilitated industry workshops (Paquin and Howard-Grenville, 2012; Van Beers et al., 2007; Mirata, 2004), IS identification systems (Grant et al., 2010), and waste exchange marketplaces (Dhanorkar et al., 2015). The key role of these methods is to facilitate information exchange of waste and resource interests (van Capelleveen et al., 2018).

Numerous pathways explain the emergence of IS, proposed in different stylized models (Chertow, 2007; Paquin and Howard-Grenville, 2012; Boons et al., 2017). Eco-industrial parks generally involve a continuous effort from coordinating bodies, e.g. municipalities or regional governments, to locate industries that can potentially cooperate together in the park regions in order to share wastes and by-products (Gibbs and Deutz, 2007). Other IS-based industrial ecosystems arose by self-organization, resulting from collaborations without top-down planning and mainly driven by economic or strategic business motivations that lead to increasing resource and waste transactions over time (Chertow and Ehrenfeld, 2012). A third type of IS emergence

is a facilitated approach that utilizes intermediaries that provide a role of strengthening trust between firms using expertise and the ability to connect industries (Paquin and Howard-Grenville, 2012). These pathways not only characterize the different types of emergence and explain how the process of IS unfolds, but also help to deduce the critical catalyzers to initiate new symbiotic actions (Boons et al., 2017).

In relation to the facilitated approach, various scholars have studied waste-exchange systems as a tool to enhance IS identification (Clayton et al., 2002; Sterr and Ott, 2004; Mirata, 2004; Chen et al., 2006; Van Beers et al., 2007; Veiga and Magrini, 2009; Grant et al., 2010; Dietrich et al., 2014; Dhanorkar et al., 2015; Cecelja et al., 2015; Hein et al., 2015; Cutaia et al., 2015; van Capelleveen et al., 2018). Such systems can enhance symbiotic transactions in the network while substantially reducing the time investment required to investigate the symbiotic potential. The capability of supporting the learning and decision-making processes of environmental problems recurs as an interesting opportunity to assess (Poch et al., 2004). Decision support tools (Grant et al., 2010) and in particular recommender support (van Capelleveen et al., 2018) are suggested as promising techniques to stimulate and facilitate the identification and assessment of new exchanges. Recommenders are able to support users in identifying item opportunities and to pro-actively engage system use, resulting in both increased sales and a more active community (Freyne et al., 2009; Pathak et al., 2010; Gomez-Uribe and Hunt, 2015). However, building systems that can provide decision support or recommend IS opportunities remains a challenge (Grant et al., 2010). Firstly, many of such tools lack the key characteristic of 'sociability' (Grant et al., 2010), focusing more on determining technical opportunities rather than building human relationships. Secondly, while a critical mass of industries is required to engage in network activity, it is difficult to attract them to join the network. Finally, systems struggle with analyzing data because of the high level of implicit knowledge, which is a burden to the development of techniques that help to identify IS exchanges (Grant et al., 2010). This particular data challenge is the key focus of this research.

For example, process data from manufacturing industries that disclose inputs, outputs, and wastes, can be both used for the identification of potential synergies as well as to provide input for recommender systems. However, the extent and level of detail with which such data is shared may be hindered

by a lack of trust among organizations because process data might reveal competitive information that organizations want to keep (partly) confidential (Paquin and Howard-Grenville, 2012). Moreover, organizations have to justify time and resource investment to explore potential ideas for which the expected benefits are not clearly predicted or even known. Therefore, the provision of detailed process data in a wider context may increase identification of IS opportunities, but this is challenging at the initial stage.

This study contributes to the discussion initiated by Grant et al. (2010) on the role of implicit and explicit knowledge in waste marketplaces by transferring the implications of data characteristics to recommendation technology. Implicit knowledge, sometimes referred to as 'tacit' knowledge, is know-how that is subconsciously understood and applied. It consists of complex information that individuals are unable to express and codify. Moreover, implicit knowledge resides primarily in its field of application. On the other side of the spectrum, explicit knowledge is formally articulated, referring to situations where one is able to capture know-how into codification schemes that can be communicated and used to reason with (Kogut and Zander, 1992; Zack, 1999). The theory on implicit versus explicit knowledge is used in organizations, supply chains, and markets to seek for explanations of problems related to communication, reasoning and broader knowledge management (Schoenherr et al., 2014; Kimble, 2013). A similar discussion on implicit versus explicit knowledge is also prevalent in the field of recommender systems (Bobadilla et al., 2013). For example, data can be derived from implicit (e.g. monitoring user's behavior) and explicit content (e.g. user ratings). Also, filtering techniques can be built using either an implicit knowledge based (e.g. association rule mining) or an explicit knowledge based (e.g. case-based reasoning) recommendation. Hence, an important question is what effects do these different recommendation techniques have on analyzing environmental (domain) data in order to create recommendations. Therefore, the aim of this study is to evaluate and understand the influential role of knowledge in the design of an effective waste material recommender for IS marketplaces.

In this paper we examine the problem of IS identification by creating a model for utilizing environmental data in recommender systems. We therefore employ the design science method (Peffers et al., 2007; Gregor and Hevner, 2013) for constructing our recommender artifact that utilizes a novel Input-Output (IO) algorithm. This IO algorithm is evaluated through a com-

4

parison with another recommendation algorithm we have developed based on Association Rule Mining (ARM). Section 2 describes the methodology used, while Section 3 presents the design of a generic approach to create explicit and implicit knowledge based recommendations in the environmental data landscape. This part provides an instantiation of our proposed model that is applied to IS by designing two algorithms, an explicit knowledge based recommender and an implicit knowledge based recommender. Section 4 presents the result of a comparison of the two algorithms. Section 5 discusses the interpretation of recommender performance and reviews the internal and external validity of the achieved results. Finally, Section 6 concludes the paper with implications and future research directions. This structure follows the guidelines suggested by Gregor and Hevner (2013).

2. Method

The main objectives of this study are to design a model that utilizes environmental data to make IS recommendations and to develop an instantiation of this model to identify IS opportunities. This design science research is guided by Peffers et al. (2007). In relation to the discussion of Grant et al. (2010), our study focuses on the role of implicit versus explicit knowledge in the design of recommenders. Thus, the instantiation of the model represents both an implicit and an explicit knowledge based recommender design. We evaluate the design empirically by investigating the extent to which explicit and implicit knowledge influences the effectiveness of the recommendations.

Our research approach consists of four major steps: (1) data collection, (2) the design of a model for utilizing environmental data to create recommendations, (3) an application of this model to the problem of identification of IS, resulting in (3a) the design of techniques for pre-processing, and (3b) the design of an implicit knowledge based as well as an explicit knowledge based recommender, and finally, (4) a comparative design evaluation using recommender evaluation metrics.

2.1. Step 1: Data collection

Two data sources are utilized as a knowledge base for making recommendations. Firstly, data is collected from industrial symbiotic workshops.

This IS data, containing a variety of waste items and resource interests from industry, originate from IS workshops held in two different European industrial clusters. The data were collected as part of the EU-funded SHARE-BOX project (European Commission, 2017). This data is explored in order to identify the item-properties that are valuable to generate useful recommendations. Secondly, knowing that confidentiality is a key challenge in IS development, we study the usefulness of an external database which provides the process input-output data from life-cycle inventories. This database contains data regarding the wastes produced and primary resources used in the production activities of process industries. In addition, it serves as a knowledge base to investigate potential synergies between industrial firms based on the substitution of traditional primary resources with wastes. A detailed overview of both data structures is presented in Section 3.2.

2.2. Step 2: Model design

After studying the data, we create a model, that uses a step-wise approach, to explain how to build a recommender system that can operate under challenging data characteristics (e.g. the noise in survey data, see Figure 1). Such noisy characteristics are often present in environmental data sets due to ill-defined and unstructured data collection practices.

2.3. Step 3: Instantiation of the model

Next, we create an instantiation of this model by applying it to the problem of IS identification based on the data gathered. According to the model (See Figure 1), firstly the data requires pre-processing before it becomes useful as input to recommender algorithms. In the case of IS data, this involves the clustering of 'ephemeral items' in the data set into latent product concepts based on a technique described in Chen and Canny (2011), which is explained in detail in Section 3.3.

Then, we design both an implicit and an explicit knowledge based recommender. Explicit knowledge refers to the use of an external knowledge base, while implicit knowledge is used for a recommender that learns from data on the behavior of users in a marketplace. The explicit knowledge based resource recommender is constructed on the concept that a waste-to-input match (i.e., a waste that can substitute a primary resource) can be predicted

6

by utilizing information about the manufacturing processes associated with a particular industry. Often, the manual use of Life-Cycle Inventory (LCI) databases containing explicit knowledge for the detection and assessment of new IS business opportunities already takes place (Grant et al., 2010; van Capelleveen et al., 2018). Industry profiles are created that contain the major inputs and outputs associated with the industry's manufacturing processes. Our explicit knowledge based recommender exploits the potential of inventory databases that provide process data revealing the inputs, outputs, and wastes associated with the manufacturing of products in a particular industry. On the other hand, the implicit knowledge based recommender utilizes the well-known techniques of association rule mining (Agrawal et al., 1993) in order to evaluate the impact of implicit knowledge on the recommendation. Association rule mining is a promising technique often applied in e-marketplaces to recommend items (Park et al., 2012). Hence, it is expected to be a good candidate for an implicit knowledge based recommendation technique.

2.4. Step 4: Design evaluation

The final step of the methodology involves running the algorithms on the IS data set and evaluating the performance of the recommenders. Usually, recommender evaluation is first performed in an off-line setting using sample data (in this case the IS workshop data). Such data is either derived from a relevant external data source or extracted from the on-line system for which the recommender is designed (Ekstrand et al., 2011). We evaluate recommendations by measuring the prediction effectiveness in terms of a binary classification. More precise, we classify the predictions as recommended items of non-recommended items and use this binary classification to measure the performance of an algorithm. Then we quantify four metrics for evaluating the performance, i.e., precision, recall, accuracy and the F-measure.

In a recommender context the precision and recall measures are described in terms of a set of retrieved items that are then compared to the set of items relevant to a user. In Equations 1, 2, and 3, tp denotes the number of true positives, fp the number of false positives, tn the number of true negatives, and fn the number of false negatives. A true positive is a recommendation

7

that matches the user's stated preference. A false positive is a recommendation that does not match the user's stated interest. A true negative regards the situation when no recommendation is provided and the user has no stated interest. A false negative occurs when no recommendation is provided to a user who has a stated interest. Precision is the fraction of recommended items that are relevant to a specific user (See Equation 1). Recall is the fraction of relevant recommended items that are retrieved (See Equation 2). Accuracy is the fraction of measurements of a true value where the true value measurements consist of both the recommended items relevant to a user and the non-relevant items that are also not recommended (see Equation 3). The F-measure is a different measure of a test's accuracy that evaluates the precision and recall in a weighted harmonic mean (see Equation 4).

$$Precision = \frac{tp}{tp + fp} \tag{1}$$

$$Recall = \frac{tp}{tp + fn} \tag{2}$$

$$Accuracy = \frac{tp + tn}{tp + tn + fp + fn} \tag{3}$$

$$F\text{-}measure = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4}$$

Recommender performance is measured at two levels, i.e. (1) an item-level, and (2) a cluster-level. This distinction is based on our assumption that many items in a resource marketplace are considered by users as similar products based on their usefulness (e.g. iron scrap from one organization is as good as those of other organizations, see also Section 3.3, where we discuss how we cluster items into latent products). Thus, we also can evaluate a recommendation as being valid if either the exact item is retrieved (item-level), or if one of the latent-product items (one of the items belonging to an item-cluster) is retrieved (cluster-level).

3. Model design: recommendation in an environmental data landscape

3.1. The model

Figure 1 presents a generic approach to create recommendations out of environmental data that is the result of ill-defined or unstructured data collection. The resulting data characteristics may not only constrain the effectiveness of a recommender but also affect necessary data transformations and taxonomy translations prior to making a recommendation. The model addresses such data issues by building recommenders based on the 'implicit' versus 'explicit' knowledge based reasoning (Grant et al., 2010; Kimble, 2013; Schoenherr et al., 2014). It contains an explanation of how implicit and explicit knowledge based recommendation algorithms are affected by the characterization of this data landscape. Explicit knowledge based recommenders require the utilization of external standardized databases containing detailed data. As explained in the model (Figure 1), environmental data is often noisy, not transparent, non-standardized, and incomplete. To create a recommendation, often a number of actions are required, including preprocessing of data, mapping data and the selection of an algorithm. Figure 1 shows that both types of algorithms (implicit or explicit knowledge based) require the pre-processing of data. But explicit knowledge based recommenders also often require the data to be mapped because of their dependence on external knowledge bases. Once the environmental data is prepared and mapped, the recommender algorithms are able to generate recommendations.
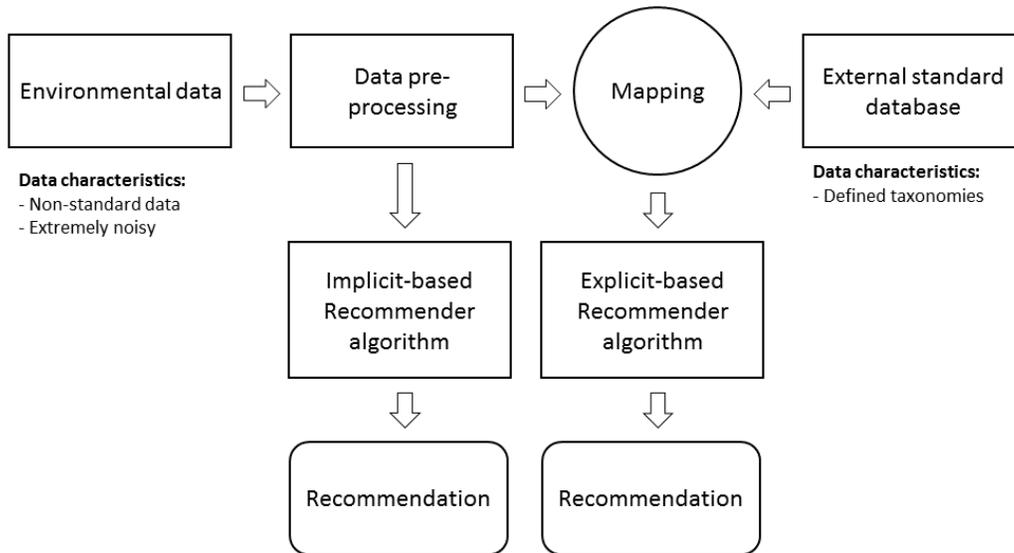
Figure 1: Model explaining the approach towards the creation of recommendation in environmental data landscapes

3.2. The model applied to the case of Industrial Symbiosis

Recommender systems can contribute to the reduction of time and resource investments of industries by guiding organizations to the first set of potential waste items that are likely to involve industrial symbiosis business cases. The following section provides illustrative examples to understand how the two recommender algorithms utilize data to construct recommendations.
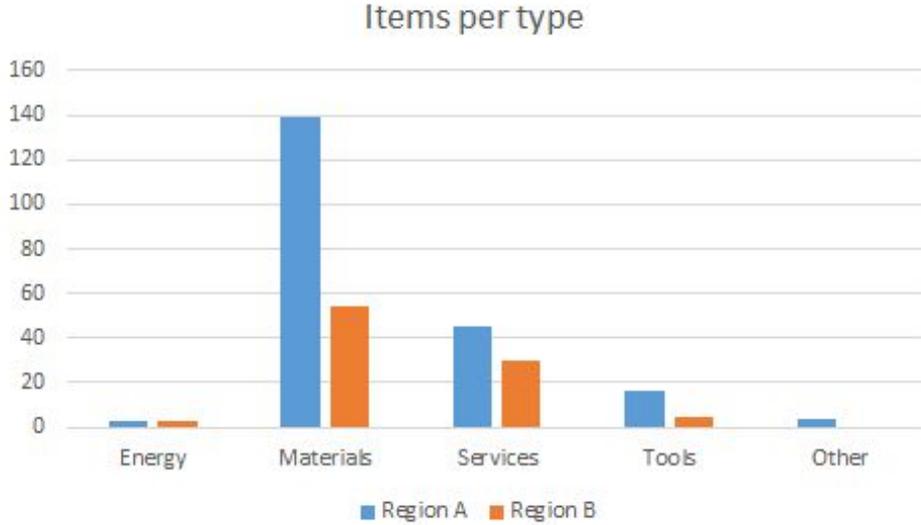
Figure 2: Items per type

Figure 2 provides information about the composition of the original workshop data. The workshops were held in two different European regions and a record was made of the waste items of participating organizations as wells as their expressed interest. This data has similar characteristics to those presented in waste ontologies for waste markets (Cecelja et al., 2015; Raafat et al., 2013). The data set has a sample size of 421 for region A, and a sample size of 150 for region B. In this data, we identify five different classifications of items, namely; (1) Materials, (2) Tools, (3) Services, (4) Energy, and (5) Others. The materials are of particular interest, as the Input-Output based recommender is designed to predict the material preference of an organization. The process industry is considered a good target audience for such a recommender. Therefore, in recommending items, organizations that do not fit the profile of a manufacturer of materials were excluded from the recommendation. Hence, in the evaluation of both recommenders, a selection of only the offered items by sellers of waste categorized as 'material' is utilized. This resulted in a total of 139 waste items for region A, and 54 waste items for region B. This sample size is relatively small compared to the evaluation of recommenders in many other types of e-marketplace studies. The reason for this is because IS waste markets do not have such large participation. Thus, careful interpretation of the statistical validity of the recommender

11

performance is needed. In particular, one needs to keep in mind that there is a possibility of overfitting the model of the ARM algorithm. In this paper, we will also derive qualitative insights on the applicability and challenges of recommender design in IS marketplaces.

Table 1: Example IS data structure

Firm	Industry[1]	Waste item	EWC [2]
A	Producer of iron castings	Iron and steel slag: Concrete tiles can be taken as one of the main components	15 01 04
B	Producer of ceramic tiles	Ceramic waste	17 01 06
C	Producer of different types of foam	Plastic bags	NULL
D	Producer of locomotives, wagons, engines	Waste oil: Oils and household waste oils in region	NULL
E	Producer of agricultural machinery	Sawmill dust and shavings	NULL
...	...	...	...

Table 1 shows a sample of the used subset of data, acquired from IS workshops organized to facilitate the creation of new IS initiatives among industries. We notice that most waste items addressed in the IS workshop survey data often have short waste descriptions, commonly defined with less than 10 words. The data illustrates the noise in item descriptions, e.g. from waste item A it is difficult to understand the exact relationship between the iron steel, the slag, and the concrete tiles. However, we can consider to a certain extent that the offered item(s) have a potential for iron reuse. The industry description contributes, as shown further on, to identify the related traditional inputs to that firm. During on-line evaluation, the ARM algo-

---

[1]Systems for statistical classifications of economic activities were beneficial to identify industry types. e.g. The International Standard Industrial Classification of All Economic Activities (ISIC) or the Statistical Classification of Economic Activities in the European Community (NACE). "NACE is derived from ISIC, in the sense that it is more detailed than ISIC. ISIC and NACE have exactly the same items at the highest levels, where NACE is more detailed at lower levels" (Eurostat, 2018).

[2]European Waste Catalogue (European Commission, 2000)

rithm can learn rules from system transactions. The workshop, however, provides off-line data containing the interests of organizations in items. The survey data is considered a representative data set for IS market transactions. In particular, because such items share the characteristic that users compose their own item descriptions when offering these in either an online market-place or during a workshop. Therefore, these stated interests are interpreted as 'transactions' to detect rules.

Table 2: Example LCI data structure (Process table)

Process ID	Process Name	
65558	cast iron production	
69729	ceramic tile production CH	
43914	polyurethane production, flexible foam RoW	
...	...	

Table 3: Example LCI data structure (Input-Output table)

Process ID	Resource ID	In/Output	Resource Name	Measure	Unit
65558	13498496	Input	Iron	8,60 E-1	Kg
65558	13498326	Input	Carbon dioxide, in air	2,51 E-2	Kg
65558	13498327	Input	Energy, gross calorific value, in biomass	2,76 E-1	MJ
65558	19624951	Output	Water/m3	1.67 E-2	M3
65558	19624487	Output	Carbon dioxide, fossil	7,15 E-1	Kg
65558	18690869	Output	Heat, waste	1,02 E-2	MJ
...	...	...	...	...	...

Tables 2 and 3 illustrate the structure of the data that is obtained from EcoInvent, one of the world's largest LCI databases. Table 2 shows examples of manufacturing processes of goods (or services) that are available, e.g. the production of iron casts. Table 3 shows examples of the inputs and outputs connected to those production processes, e.g. iron. Furthermore, information

is included regarding the amounts usually associated with the production of one functional unit.

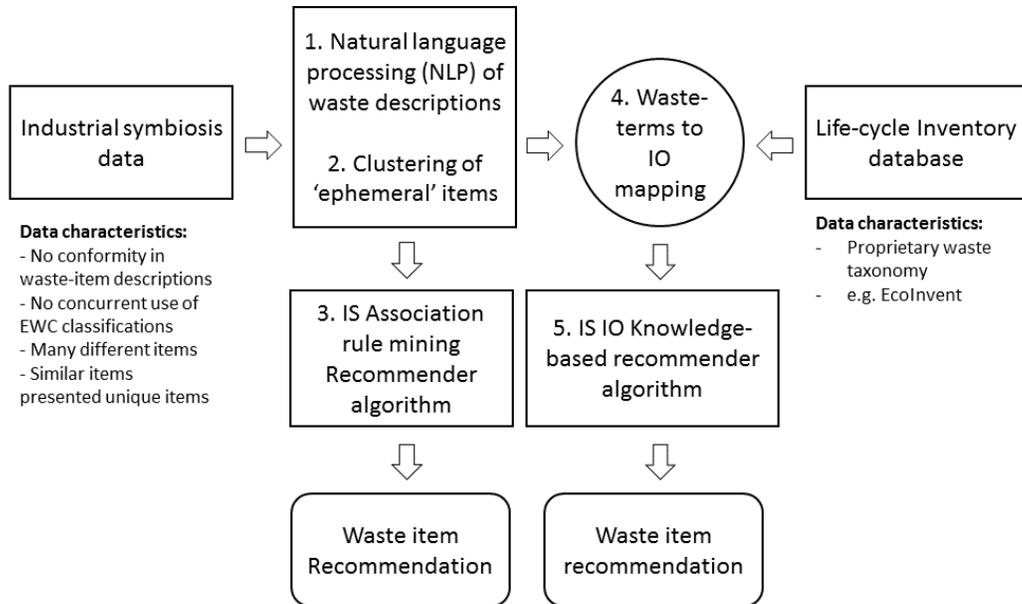## Model application to an Industrial symbiosis IO-recommender

Figure 3: Model applied as industrial symbiosis recommender

Using the analysis of the IS and LCI data, we can build an instance of our earlier proposed model (cf. Figure 1) that addresses the challenge of identifying new IS business. Figure 3 illustrates this instantiation. In Figure 3, the IS workshop data represents the 'Environmental data' box from Figure 1 and the LCI database represents an 'External standard database'. The particular data characteristics represented in Figure 3 help in the design of the data pre-processing techniques, the mapping, and the recommender algorithms. In the following three sub-sections (Sections 3.3, 3.4, and 3.5), we explain how we handle the pre-processing of the data, i.e. the need to apply Natural Language Processing (NLP) in order to cluster groups of similar items (Steps 1 and 2 in Figure 3), after which we show the mapping of items on the input-output database and the design of a recommender algorithm based on this explicit input-output knowledge (Steps 4 and 5 in Figure 3), as well as the design of a recommender algorithm based on implicit knowledge, using association rule mining (Step 3 in Figure 3).

14

3.3. Pre-processing IS data

The pre-processing of data is addressed by Steps 1 and 2 in the model in Figure 3. This consists of applying NLP and of the clustering of items. A major challenge to many recommenders is that data are often sparse (e.g. due to low item transaction history), and hence item space reduction is needed, similar to data in e-commerce marketplaces that predominantly consist of 'ephemeral items'. These 'ephemeral items' are items submitted by users and composed of users' individual product descriptions. Thus, item descriptions hardly correspond to any catalog taxonomy and often lack detailed product descriptions. In addition, the frequency with which the supply of items is renewed in a marketplace is high (Wroblewska et al., 2016). Therefore, item space reduction strategies are often applied to reduce the number of item types, increasing the number of transactions of each item type (e.g. considering all items listing iron fillings as equal iron waste products). Item space reduction methods thus create a richer history for each item type by grouping items and treating them as one 'item space', thereby allowing algorithms such as ARM to deduce more and stronger associations. Moreover, it can improve the IO matching of items with short item-descriptions by using a richer product concept description. Hence, grouping similar items helps to increase quality, efficiency, and effectiveness of recommender algorithms (Chen and Canny, 2011). NLP is used to assess the similarity between waste item descriptions.

An obvious choice to group waste market items would be to rely on existing taxonomies that already play an important role in waste treatment. For example, the European Waste Catalogue (EWC) (European Commission, 2000) or the Central Product Classification (CPC) (United Nations Statistics Division, 2015) are examples used in reports on waste disposal and in 'duty of care' documents in waste transfers (Natural Resources Wales et al., 2015) that define the similarity between items. However, such classifications are often absent in item descriptions. Moreover, these systems classify goods and services in the industry from which they originate, causing an overlap in product concepts within the taxonomy. Thus, such a taxonomy fails to relate two similar waste items if they are produced in different industries (The ISDATA project, 2015; Sander et al., 2008). For example, recycled glass can be produced either with uncontaminated glass residues resulting from a glass

bottle production facility, or it may be extracted from construction and demolition waste (International Synergies Ltd., 2016).

A solution to this problem is to map items into 'latent product concepts' (Chen and Canny, 2011). By inferring the 'latent product concept', the dynamics and diverse item-inventories can be used to group items considered as similar or identical products by a recommender. This way, item-clustering reduces complexity in an intermediate step towards enabling recommender algorithms to learn from data in such contexts. Hence, there is a trade-off between having item-specific information to be used by a recommender (enable individual based reasoning) and building item history on similar types of items (history is required to deduce user preference). The 'elbow' technique (Salvador and Chan, 2004) is used to detect the cut-off point for clustering thereby identifying the trade-off.

Items:
1. Iron and steel slags
2. Iron fillings
3. Sawmill dust and shavings

$$\begin{array}{ccccccc} \text{iron} & \text{steel} & \text{slag} & \text{fill} & \text{sawmill} & \text{dust} & \text{shave} \\ 1 & 1 & 1 & 0 & 0 & 0 & 0 \end{array}$$

Figure 4: Stem frequency item vector for item 1, constructed from item 1, 2 and 3

The clustering algorithm (See Appendix A.1), based on (Chen and Canny, 2011), explains how to cluster similar waste items. The algorithm extracts the 'latent product concept' from item descriptions in the IS data set. It uses stem-frequency vectorization as a means to identify the 'latent product concept' of an offered waste item. Stemming is the process of removing the inflectional forms and sometimes the derivations of the word, by identifying the morphological root of a word (Manning et al., 2008). The frequency vector of an item within a set of item descriptions is determined by the frequency of every unique stem in that item along with the unique stems in the set. An example of such an item-vector is provided in Figure 4. In order to derive the set of stems from an item description, we use the NLTK package, a platform for working with human language data in Python (Natural Language Toolkit, 2017). Firstly, all the characters are converted to lowercase and all numbers and special characters are removed. Then, items are tokenized into a bag of words. Using the NLTK corpus, all English stop words are removed

from these bags and some non-significant terminology commonly used in IS is filtered as well, e.g. 'waste', 'material', and 'process'. Finally, the empirically justified Porter algorithm (Porter, 1980) is applied to stem the bag of words. Then, the resulting item vectors are utilized in a multi-dimensional hierarchical agglomerative clustering algorithm, based on an algorithm presented in Manning et al. (2008), to cluster items using the cosine similarity of item vectors (See Appendix A.2). In contrast to most clustering methods, in our multidimensional clustering, the traditional coordinates x,y are replaced by the stem-frequency vector.

In hierarchical clustering, the determination of the 'elbow' is a common method to determine the number of clusters. The method utilizes statistical variance in order to explain the best number of clusters. The variance is calculated using the Sum of Squared Errors (SSE) as a function of the decrease in the number of clusters. The method exploits the point of maximum curvature in the Sum of Squared Errors (SSE) of all clusters as the cut-off point for clustering. The maximum curvature is the point on the curve with maximal bending, also explained as the point where the curve appears to "curve" the most (Salvador and Chan, 2004). This point often represents a 'knee' or 'elbow', i.e., the point that shows a marginal drop or gain in variance. This point determines the finally selected number of clusters. Although evaluation methods that can mathematically locate the 'elbow' in a curve (Salvador and Chan, 2004) exist, visual interpretation is preferred over statistical methods when no objective measures have been defined for cluster optimality in the domain of application (Jung et al., 2003).

3.4. Design of an explicit knowledge based IS recommendation

This section addresses Steps 4 and 5 in Figure 3, i.e. the mapping of items to the input-output data in the LCI database and the IO algorithm design. The IO based recommender (See Algorithm 1) is a knowledge based recommender, suggesting items based on the inferences about the needs and preferences of a user (Burke, 2002). This type of recommender makes use of explicit knowledge (a knowledge base) to recommend items to users. The algorithm is based on the assumption that the items offered in IS marketplaces correspond to the specific item interest of an organization, and subsequently to the primary inputs in their manufacturing processes. As primary inputs of a production process, we consider only those inputs that are not provided

by one of the production processes of the associated production chain, such as raw materials and natural resources (e.g. silver ore, limestone, clay, water) and energy resources (e.g. gasoline, natural gas, coal). The primary outputs result from a manufacturing process, which is the primary focus of production. Secondary output generally refers to waste that has no perceived economic value and that traditionally is to be discarded. As this method uses the identification of potential relations between the primary inputs of production processes and secondary output waste offered in a marketplace, we indicate this type of raw material recommender as an input-output recommender (IO recommender).

The information on resource use for the production of goods or services originates from an LCI database, in our case the EcoInvent database, version 3.3 (Ecoinvent, 2017). This type of database provides well-documented process data for thousands of production processes, including information on the raw material consumption. In particular, this information about primary inputs is used to identify the most consumed raw materials of an organization and matches those to the potential waste resources available in the marketplace.

Recommendations made by the IO recommender algorithm are created through the use of pre-compiled manufacturing profiles. A profile is constructed through the identification of the type of products produced by an industry, typically constructed using the corresponding company websites. The products are used to find the associated manufacturing processes available in LCI databases. This results in predefined profiles created for each type of organization (e.g. a manufacturer of castings). The profiles list, among others, the raw material consumption that is associated with the selected manufacturing processes. The profiles of the industries required by the algorithm are constructed prior to recommendation. Once the profiles are created, the algorithm extracts all resources listed from the LCI database that is linked to a production process in the organizational profile. Then, it filters the resources which make up the largest fraction of resource consumption within the production of a product. These resources are selected as the candidates for recommendation. The resource candidates are then compared with the marketplace item descriptions to see whether they contain similarities. The matching algorithm calculates the cosine similarity between the vectorized stem-frequency of a waste product description (a cluster) and

that of a resource description. The stem-frequency vectorization algorithm applied in the clustering of items (See Section 3.3) is identical to the one used for IO matching.

The algorithm operates as follows. First, for each resource connected to the industrial profile of the organization, the algorithm iterates over all clusters in the marketplace. Within this iteration, a bag of stems is created for both the selected cluster (the latent product concept) and the selected resource (from the industry profile). The first bag of stems, created from the cluster, selects the most frequently occurring stems in all items (descriptions) belonging to that cluster. The second bag of stems is created from the resource and uses the name of the material based on the taxonomy used in the LCI. These bags of stems are used to compose the item vectors for both the cluster and the resource. Next, the similarity between the two item vectors is assessed using a cosine similarity function. The stem frequency item vector combinations that exceed the minimum cosine threshold and are not yet a recommendation, are added to the set of recommendations.

Algorithm 1: Input-output recommender algorithm

Data:

$N_{resources}$ = Set of associated resources of an organizational manufacturing profile derived from LCI

W_{items} = Set of all waste items in marketplace

Parameters:

min_cos = Minimum cosine similarity

Variables :

B = Bag of stems

V = Stem-frequency item vector

Result:

R = Set of recommendations

1 **Function** *Input-Output-recommender($N_{resources}, W_{items}, min\_cos$)*

2   **foreach** *resource* $\in N_{resources}$ **do**

3   **foreach** *cluster* $\in W_{items}$ **do**

4   Create a bag of stems $B_{cluster}$ with the n most frequent stems from all items in the *cluster* where n is the average number of stems in each item belonging to that cluster;

5   Create a bag of stems $B_{resource}$ from the item *resource*;

6   Create a stem-frequency vector $V_{cluster}$ for the *cluster* with $B_{cluster}$ as item;

7   Create a stem-frequency vector $V_{resource}$ for the *resource* with $B_{resource}$ as item;

8   Calculate the cosines similarity s between cluster vector $V_{cluster}$ and resource vector $V_{resource}$;

9   **if** $s > min\_cos$ and *cluster* $\notin R$ **then**

10   Add a unique recommendation *cluster* to R;

11   **return** R;

3.5. Design of an implicit knowledge based IS recommendation

Association rule mining is one of the techniques based on implicit knowledge that found successful application in recommender systems (Park et al., 2012). The ARM technique attempts to discover regularities in transaction data based on the concept of strong rules (Agrawal et al., 1993). In general, association rules perform best on large-scale data with a broad history of transactions. The popular Apriori algorithm is used to generate candi-

dates for identifying these rules, as it is both simple and exact (Agrawal and Srikant, 1994).

The pace at which users change preference affects the decay of association rules. To find the best length of the time period from which rules can be detected requires an estimate of how stable the user preference is over time. The nature of IS is characterized by a low frequency of transactions, but preference is considered to remain fairly stable. Therefore, in the context of IS, it is reasonable to select a longer period from which association rules may be learned. This period ranges from several months up to years.

To recommend on the basis of association rules, one needs to define which rules are accepted to generate recommendations. It is common to set a threshold for the minimal support, and the minimal confidence in order for association rules to become a rule for recommendation. Support is a measure of item pair frequency. It is defined as the percentage of transactions in a dataset that contains a particular item combination. Confidence is a measure of rule strength, defined as the number of times an association rule is valid. In case of a relatively small dataset k-fold cross-validation is applied in order to increase the validity of the evaluation. K-fold cross-validation is a technique that randomly partitions data into equally sized samples in order to test an algorithm in various test and training combinations.

Algorithm 2: (k-fold) association rule mining recommender

Data:
W_{items} = Set of all waste items in marketplace

Parameters:
min_sup = Minimum support
min_conf = Minimum confidence
k_value = k-value of k-fold validation

Variables :
$I_{training}$ = List of training samples
C = Candidate rules

Result:
R = Recommendations for each train/test sample
I_{test} = List of test samples

1 **Function** $k\text{-}fold\text{-}ARM(W_{items}, min\_sup, min\_conf, k\_value)$
2   **for** $i \leftarrow 1$ **to** k_value **do**
3   Divide W_{items} into k_value equally sized samples s;
4   Assign all samples s, except where $s = i$, to training set $I_{training}$;
5   Assign sample s where $s = i$ to test set I_{test};

6   **foreach** sample $s \in I_{traning}$ **do**
7   Calculate the candidates $C_{candidates}$ and the support $C_{support}$ from the sample s with a minimum support of min_sup;
8   Generate the association rules r from the candidates $C_{candidates}$ and support data $C_{support}$ with a minimum confidence min_conf;
9   Add the rules r to the rules;

10   **return** R, I_{test};

---

4. Design evaluation

4.1. Data preparation

Figure 5 presents the results of how clusters were formed using the maximum curvature concept (Explained in Section 3.3). The resulting number of clusters k is 84 for region A and 40 for region B respectively. The clustering is applied to both data sets separately. The remaining parameters of the algorithms are consistently used and initialized with similar values for both data sets.
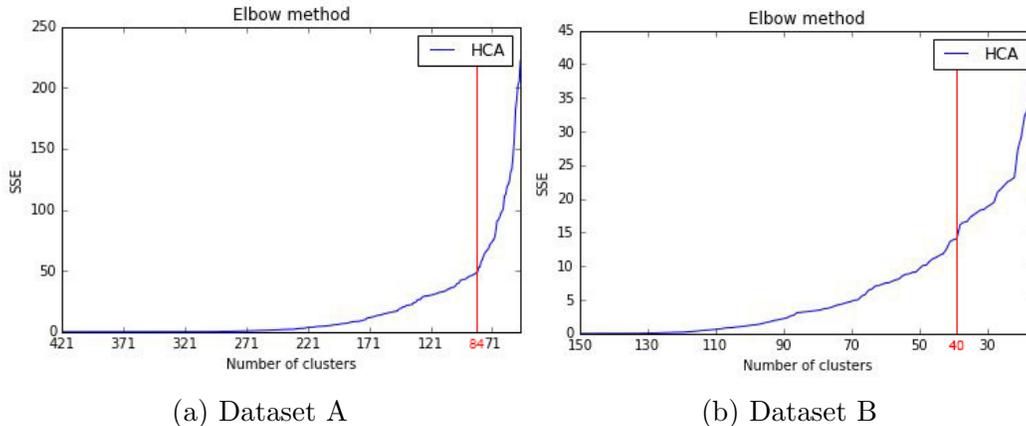
(a) Dataset A   (b) Dataset B

Figure 5: The Sum of Squared Errors (SSE) is plotted to detect the cut-off criteria for clustering.

For the input-output algorithm, a minimal cosine similarity of 0.1 is set, reflecting that one of the stems derived from the item description matches the resource taxonomy of the LCI. This value needs to be kept low, as nearly all item descriptions contain more stems (often not more than 10) than the solely resource taxonomic term used in the LCI. For example, a resource taxonomy names a material 'Iron', while the item description contains more words than the resource term only, e.g. "Iron steel and slag: concrete tiles can be taken as one of the main components" (see Table 1). Of course, having more stems to consider increases the likelihood to find a match between the IS data and the external database. However, it may also increase the number of incorrect matchings.

The k-fold cross-validation technique used in the ARM uses the value $k = 10$. Furthermore, the minimal support parameter is initialized with a value of 0.1 and the minimal confidence parameter with a value of 0.7 as thresholds for a recommendation rule. The value for minimal confidence is selected based on a sensitivity analysis which examines the highest potentially achievable result (Martnez-Ballesteros et al., 2016). Figure 6 shows this sensitivity analysis from which we obtained the optimal initialization value of minimal confidence. The intersection of lines representing precision and recall is generally recognized as a good balance (see the crossing of lines in Figure 6f). Although it could be desirable in a system implementation to exploit precision in favor of recall and a larger number of recommendations,

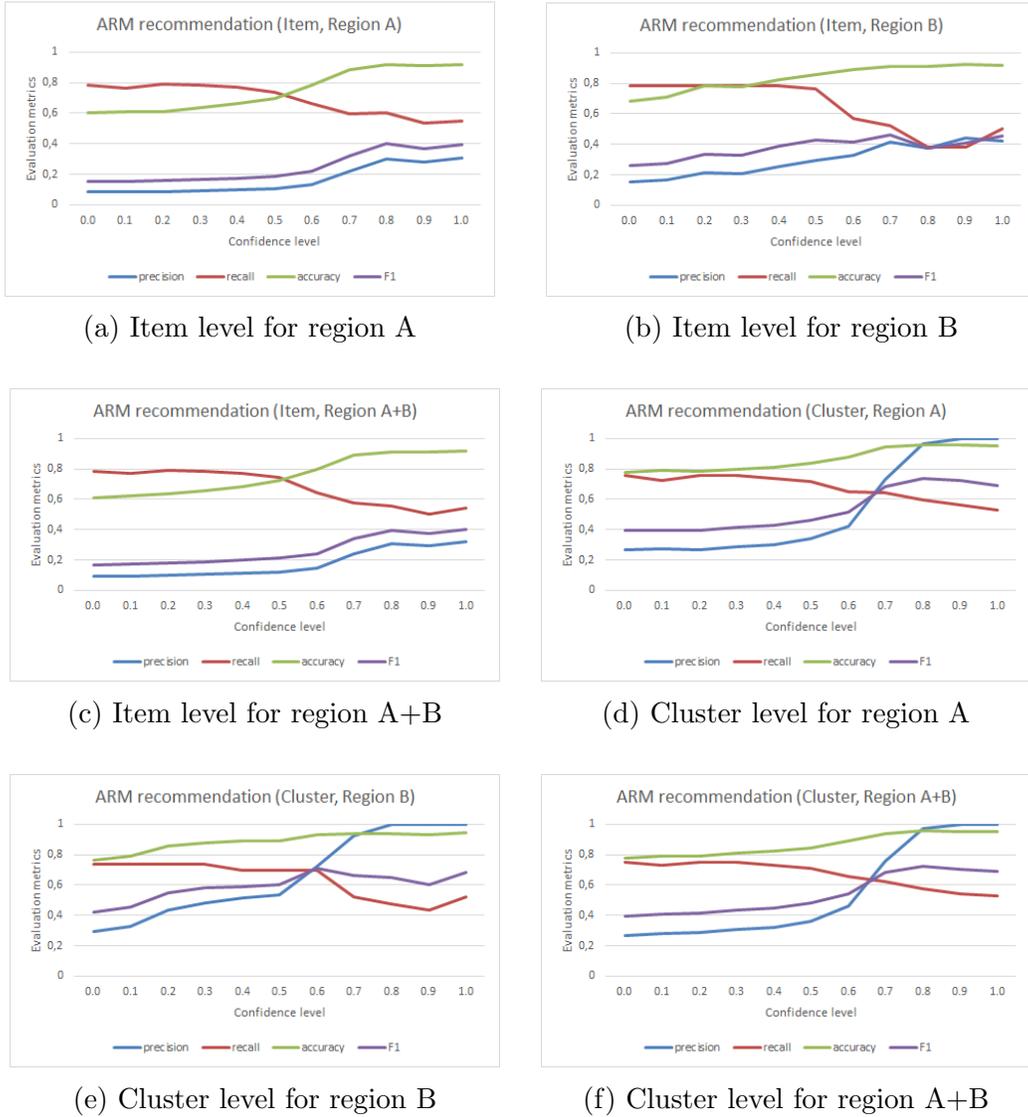we select this point to present the lower bound for our recommender system.

(a) Item level for region A

(b) Item level for region B

(c) Item level for region A+B

(d) Cluster level for region A

(e) Cluster level for region B

(f) Cluster level for region A+B

Figure 6: Sensitivity analysis of confidence level in ARM

4.2. Results of recommender performance

Table 4 shows the results of a comparison of the two recommenders. The results are presented with three variables. In the first column, we distinguish

between the two recommender methods that were tested. The second column shows the level of measurement, item-level or cluster-level, as explained in Section 2. The third column defines the data set used to test the algorithm. Note that the average performance for the combined regions (A+B) is constructed by first aggregating the recommendation results of the individual regional experiments in order to calculate the combined performance metrics. Table 4 also presents the percentage of items that were recommended, as well as the values of all performance metrics: precision, recall, accuracy and the F1 or F-measure.

Table 4: A comparison of recommender performance

Region	Level	Method	Recom. items	Precision	Recall	Accuracy	F1
A	Cluster	IO	0.1417	0.1210	0.1792	0.7969	0.1444
A	Cluster	ARM	0.0839	0.7312	0.6415	0.9431	0.6834
A	Item	IO	0.1706	0.0564	0.2061	0.8020	0.0885
A	Item	ARM	0.1245	0.2227	0.5939	0.8843	0.3240
B	Cluster	IO	0.1717	0.1764	0.2608	0.7727	0.2105
B	Cluster	ARM	0.0657	0.9231	0.5217	0.9394	0.6667
B	Item	IO	0.2570	0.0516	0.1904	0.7000	0.0812
B	Item	ARM	0.0879	0.4151	0.5238	0.9154	0.4632
A+B	Cluster	IO	0.1462	0.1309	0.1938	0.7932	0.1562
A+B	Cluster	ARM	0.812	0.7547	0.6202	0.9426	0.6809
A+B	Item	IO	0.1832	0.0554	0.2029	0.7871	0.0870
A+B	Item	ARM	0.1191	0.2434	0.5797	0.8888	0.3429

The IO and ARM method are first compared at cluster level. The results demonstrate that the ARM method outperforms the IO method on average with a substantial difference in precision, ranging from 4 to 6 times, and 2 or 3 times on recall. In general, also the evaluation of the accuracy and F-measure shows a noticeable difference in performance. A similar substantial difference is found at the item level. The results of the ARM method compared to the IO method is 3 to 4 times higher for precision and 2 to 3 times for recall. That both methods score much lower at the item level is most likely explained by the number of similar items organizations purchase. For example, if three similar items, grouped into a latent product, are available

and one organization only purchases two of these, then the precision at item level is lower than at the cluster level.

The scores on accuracy require more detailed explanation. The results show that the accuracy values in the experiments are less different, ranging from 0.7 to 0.9. However, accuracy measures may portray a misleading perspective of recommender performance. Although it is true that the accuracy results indicate that many items were correctly classified, the majority of correct classifications in this experiment consists of items that were correctly rejected for recommendation. The effectiveness of a recommender is more clearly understood by having high precision rates along with a reasonable recall, rather than by the overall accuracy result. Therefore, also the F-measure is calculated. F-measure provides a combined perspective on precision and recall and a better demonstration of the difference in performance of the recommenders.

Finally, the two regions show a small difference in the performance of recommendation algorithms. Regional characteristics may have affected the performance of the algorithms. The IO method shows only small differences between the regions. However, the ARM method shows a larger disparity between regions, i.e., Region B performs considerably better at precision than Region A, while no such precision increase is achieved with the IO method. This can be explained by the fact that in a relatively small data set only a small number of rules can be deduced. It can be understood that in region B the recommender model overfits based on the few learned rules. With a larger dataset, it is likely that more rules would be available, leading to a considerable drop in confidence levels. Consequently, a lower precision-recall balance would be established based on the larger set of rules.

The Wilcoxon signed-rank test is applied to show that the results obtained in our experiment are significantly different (See Table 5). This test compares the two algorithms based on the paired results (between industries) without making distributional assumptions on the differences (Shani and Gunawardana, 2011). Hence, the test that detects whether the performance difference between the ARM and the IO recommender is statistically significant when evaluating at the cluster level using data from both regions A and B. Such a hypothesis is non-directional and therefore the two-sided test is applied. All tests are applied at a 95% confidence interval, thus using an α of 0.05.

The resulting Wilcoxon critical T values for each sample size n are obtained from the Wilcoxon critical T values list (Wilcoxon and Wilcox, 1964). The selected Wilcoxon test adjusts for the ranking in case of ties. Furthermore, it is initialized with the setting that accounts for the likely binomial distribution in a small data set by applying continuity correction. Both the precision and recall results are significant at p $\leq$ 0.05. Because $n \leq 20$, we confirm the difference by showing that T $\leq$ 52 for precision and T $\leq$ 46 for recall. Tests at item-level also result in a confirmation that there is a significant difference for both precision and recall between the ARM-method and the IO-method (See Table 5).

Table 5: Wilcoxon signed rank test

Test	Test-data	Metric	n	Critical T	T	α	p
Wilcoxon two-sided	Cluster-level (Region A+B)	Precision	20	52	4.5	0.05	0.00019
Wilcoxon two-sided	Cluster-level (Region A+B)	Recall	17	46	16.5	0.05	0.00481
Wilcoxon two-sided	Item-level (Region A+B)	Precision	15	25	2	0.05	0.00109
Wilcoxon two-sided	Item-level (Region A+B)	Recall	18	40	18	0.05	0.00352

5. Discussion

5.1. The role of explicit and implicit knowledge

For the total set of available items, the experiments show that a substantial number of recommendations can be generated using both the ARM and the IO-based methods. However, the results indicate a preference to utilize implicit knowledge over explicit knowledge in recommender design for IS identification. In both regional data sets measured at item and cluster levels, the ARM algorithm is clearly the better performer for most of the recommender metrics, except for accuracy that shows a weaker and less consistent difference. These findings confirm the ideas of Grant et al. (2010) that implicit knowledge is a challenging characteristic of industrial symbiotic markets that partly affects the success of decision support designs. Nevertheless, we see a role for recommenders in IS markets, in particular for implicit knowledge based algorithms. With high precision rates, recommenders

could attract organizations to form a critical mass that can sustain sufficient supply-demand in the marketplace. Such recommenders could be applied through e.g. newsletters, that are used as a recurring tool by organizations to be notified of new items after they have participated in workshop sessions. In such a way, slowly we may grow from separate business transactions at workshops towards an active private e-marketplace.

Balancing precision and recall is a frequent concern of many recommender system designers, e.g. see Geyer-Schulz and Hahsler (2003). In general, we target customers in marketplaces with a preference for precision over recall providing good recommendations irrespective of the number of recommendations (which is what precision measures). In case users don't mind seeing a few extra irrelevant recommendations. In order to receive more recommendations they would be interested in, we could optimize on recall. However, it is often the case that over time, optimizing on precision may limit the users' exposure to novel item types, whereas one of the goals of an IS recommender is also to initiate new types of symbiotic ideas. A trade-off mechanism between all such performance measures (e.g. accuracy, novelty, dispersity, and stability) with respect to the delivery format (predictions vs. top-n), can provide a more balanced evaluation of recommender goals (Bobadilla et al., 2013). Therefore, the performance of a recommender is always highly suggested to be frequently evaluated within the context of the application, according to the changing preferences of its users.

The process of industrial symbiotic business development is generally categorized in four phases: identification, assessment, implementation, and monitoring. Recommenders proposed in this paper address the first phase, i.e. identification. Therefore, the initial baseline for using recommenders for more complex domains is that the success rate of recommendation is at least considerably better than random recommendation or search ('information filtering') approaches. Our results are not comparable to those obtained in B2C markets where the precision of e.g. movie recommendation is more likely to be high (most consumers like many movies) compared to the precision in B2B markets characterized by narrow interests. Our expectation is that ARM could perform better in specialized markets that reflect a narrow business interest. Current IS markets offer all sorts of items varying from waste heat to iron scrap metals. Specialized markets, such as metal markets, have fewer sorts of items. This potentially helps recommenders because of

the expected stronger rules that can be deduced. However, once organizations start to use the recommenders for identifying business opportunities, they might improve the precision and recall based on their business experience in the successive phase of the assessment. To begin with, organizations may already save costs when recommenders are able to substantially reduce the number of irrelevant matches.

5.2. Data challenges to explicit knowledge based IS recommenders

The IO recommender was able to deduce item preference given the fact that preferred items were retrieved. This demonstrates the value of studying explicit knowledge based recommendation. However, it is questionable whether the precision of the IO algorithm can be improved to become an effective predictor in an industrial symbiotic marketplace. For this, data quality and standardization issues first need to be improved in order to determine whether this form of explicit based recommendation can also perform at the rate ARM is currently able to. Under the conditions of the current IS data landscape, we do not expect to reach such results anytime soon.

Data quality has always been considered as one of the major challenges that need to be addressed by organizations working on predictive analytics projects in supply chain management (Hazen et al., 2014). To understand the problem of data quality in recommendations, a number of data quality dimensions need to be assessed (e.g. accuracy, timeliness, consistency, and completeness). Furthermore, ontological problems that have an impact on the effectiveness of prediction or matchmaking technologies have to be taken into account (Cecelja et al., 2015). Current decision support tools operate with data from industrial symbiotic markets gathered under real-world business conditions. As a result, the level of detail and information richness is generally limited. In addition, approximate estimates are used in quantitative values and often a variety of data formats is used, leading to unstructured data. Poor data quality not only hinders the determination of an IS match but also hinders better identification of potential opportunities.

Table 6: Data challenges to input-output based industrial symbiosis identification

Data challenge	Severity	Literature support
- Use of different waste or material taxonomies	High	
- Addressing waste at multiple nodes or different levels in the hierarchy of a taxonomy	High	Overlap in taxonomy (Sander et al., 2008)
- Limited detail in waste descriptions	High	Waste-ontology Cecelja et al. (2015)
- Limited structured data available within the public domain on material alternatives to identify substitution	High	Substitution alternatives (Hein et al., 2015)
- Derive a decomposition of materials from waste descriptions to reveal individual opportunities	Mid	Material decomposition (Yuan et al., 2013)
- The used process inputs data may not directly reflect the actual industry interests	Mid	
- Production processes data used to compose the organizational profile do not perfectly reflect the actual production process of the involved organization	Mid	
- Process input data from LCI databases are inaccurate or outdated	Low	
- Process data from either organizations or LCI databases can be unreliable with respect to data accuracy, timeliness, trustworthiness, consistency.	Low	Uncertainty of production volumes (Leong et al., 2016), Missing product and transaction information (Dhanorkar et al., 2015), Trustworthiness of data (Sheng et al., 2005)
- Process data is limitedly quantitative (related to production frequency, production volumes and waste-quality)	Low	

During the design of the IO recommender, a number of data quality prob-

lems were encountered. Table 6 illustrates the reasons for these knowledge mismatches that result in bad or missing item-recommendations. We classify the data challenges with a severity level based on the expected frequency of each type of challenge in the IS data. A number of studies illustrate these data challenges, which we found to obstruct knowledge based recommendation (see column 'Literature support' in Table 6). For example, a waste offer that lists certain types of bio-materials that can be used to produce bio-energy may not be directly linked to a demand of bio-energy. In order to let a system suggest a logical knowledge based recommendation, the system should be aware of the link between bio-materials as a resource to produce bio-energy or should be able to infer the relation between the offered waste and the demand for bio-energy, e.g. based on linked or historical data. One of the major data concerns is encountered in the different hierarchical levels in which wastes are addressed (Sander et al., 2008). Consider the relationship between 'iron' and 'metal'. Iron refers to raw material largely consisting of the chemical element Ferrum, is a type of metal and likely to be found as scrap metal. However, without such an explicit relation, text-mining algorithms would struggle to relate the two.

Based on this analysis, there are three main strategies to make recommendation work in an IS context. (1) Increase the data quality, so that knowledge based recommenders have better data to reason with and therefore are likely to make better predictions. (2) Adapt the knowledge based recommenders to work with limited and less accurate data. Evaluate if the recommenders can perform at an acceptable rate of recommendation. (3) Rely on implicit knowledge based recommenders that are not dependent on the quality of data in external knowledge bases.

5.3. Internal and external validity

Considering the internal validity of the study, several remarks are in place. The evaluation indicates that in a practical IS context an algorithm based on implicit knowledge perform better in predicting item preference than those based on explicit knowledge. This finding is in line with the study of Grant et al. (2010) and indicates that the theory explaining the role of implicit versus explicit knowledge is transferable to the design of IS recommenders. However, the study selects and designs two algorithms based on existing concepts of recommender systems related to implicit and explicit knowledge (e.g. ontology based versus association rule based). Therefore, the evaluation of

the design, a part of design science methodology, relies on qualitative inter-pretation. A more in-depth analysis might include other filtering techniques that represent the implicit and explicit knowledge based recommendation to strengthen the validity of our research. However, based on previous litera-ture (Grant et al., 2010), we believe that different performance outcomes are highly unlikely because of the typical data problems faced in IS marketplaces.

Another aspect of internal validity is that the LCI database used to create company profiles has its limitations. Not all industrial manufacturers' pro-cesses could be identified, thereby resulting in fewer recommendations. With respect to ARM, training and testing the algorithms on a larger data set might improve the results. Hence, the ARM is not considered to be heavily influenced by potentially negative effects, e.g. by overfitting the classifier on this IS data set. Finally, we focused on the exchange of materials, i.e. waste items and process industry organizations were selected to study the role of implicit and explicit knowledge in the design of an IS recommender. IS is a wider concept that goes beyond the exchange of raw materials. The study did not enclose IS item types such as waste energy, tools & machinery or over-capacity in a service-based industry.

Regarding the external validity, a number of generalizations can be de-rived from this research. Firstly, the proposed model presented in Figure 1 is applied to other cases in which recommenders are designed using en-vironmental data that are extracted from ill-defined or unstructured data collections. The approach can be applied to all types of environmental prob-lems characterized by such typical survey data that one might use together with external data to generate recommendations. Secondly, the instantia-tion of the model for the case of IS (Figure 3) could be generalized to other cases of recommender design for similar symbiosis settings. We believe that the data structures of IS and LCI are representative for other situations in which IS identification is sought. As this study is a first attempt of building a recommender in an IS context, we do not believe these designs are exclu-sive. On the contrary, although these designs may be transferable to other IS settings, other types of implicit and explicit knowledge based recommender filtering techniques (e.g. rule-based) may replace them and show a different performance.

The generalizability of the role of knowledge in the recommendation is

affected by the quality of available data sources. For example, EcoInvent as a knowledge base of input-output data is currently the best possible source for an IO-based recommender. Although we expect the quality of such sources to increase, they require substantial collaborative efforts that can be part of long-term development goals. Moreover, the proposed model can be applied under circumstances having such data quality issues. Our methodology explicitly attempts to work with these data sets without considering the possibility of iterative refinement of data quality in collaboration with users. The reason for this approach is that if extensive efforts are associated with improving the data quality, industries are unlikely to make the necessary investments. This has a more direct impact on the sustainability of similar industries.

6. Conclusion

A recurring type of IS research is the development of software applications to support the process of facilitating symbiotic developments. This paper addresses one of the major problems for the adoption of IS identification tools in a practical context. Following the theory of Grant et al. (2010) on the dominance of implicit knowledge in IS markets, we test the role of implicit versus explicit knowledge in designing IS recommenders. We find that the implicit knowledge based recommender significantly outperforms the explicit knowledge based IS recommender. In other words, the performance of the IO recommender is considerably lower than the ARM recommender algorithm that relies on implicit knowledge. The study further shows that the design of an explicit knowledge based recommender is affected by many data challenges, including the linkage of waste streams to process inputs, structural and semantic representation of data, attribute availability, code standardization, data reliability and data integrity problems. The design evaluation shows that implicit knowledge based algorithm can outperform explicit knowledge based algorithms in identifying IS opportunities. We argue that such a performance difference can be explained by a number of data challenges that first have to be resolved before explicit knowledge based recommendation can be of practical value to IS decision support. On the other hand, we observe a noticeable role for implicit knowledge based techniques in recommendation. The current performance of implicit knowledge based recommenders might enable its acceptance among industries that intend to investigate potential IS opportunities. This provides an indication to practi-

tioners that usually implicit knowledge based algorithms are more promising for the design of IS recommenders. On the other hand, a new research challenge emerges on how to improve the environmental data landscape so that explicit knowledge based recommendation can become a more viable option. Finally, the results clearly demonstrate the benefits and limitations of both the IO and the ARM method. More generally, they demonstrate the challenges and preconditions for a successful recommender design in IS markets. Moreover, we confirm that recommender evaluation is essential to the design of an effective prediction algorithm.

The following contributions are provided by this study. (i) To the best of the authors' knowledge, this paper is a first attempt in the IS literature to propose a recommender for IS waste markets and to test the proposed recommenders with real-world data. (ii) Our proposed model explains the design of recommenders in an environmental data context subject to data quality issues. (iii) In the instantiation of the model, we construct two algorithm designs based on explicit and implicit knowledge. The proposed explicit knowledge based recommender algorithm uses manufacturing profiles created with input-output data of life-cycle inventory databases. Such a recommender demonstrates the applicability of knowledge based recommendations. (iv) We provide comparative performance results for both recommenders so that practical implications are provided based on statistically tested performance indicators.

This first attempt to implement recommenders in IS provides insights on how to design recommenders in symbiotic networks. In future work, we intend to address other types of recommenders. Our findings, along with the support of Grant et al. (2010), indicate that given the current issues with IS data, one might preferably focus efforts and resources in designing recommenders based on implicit knowledge rather than explicit knowledge. IS recommender design can also be studied in a broader perspective of the identified IS categories, including waste energy, services and manufacturing tools. In addition, different recommender techniques can be discussed, such as IS sector-based recommendation, along with further testing various recommendation mechanisms based on implicit and explicit knowledge.

7. Acknowledgements

References

Agrawal, R., Imieliński, T., Swami, A., Jun. 1993. Mining association rules between sets of items in large databases. SIGMOD Rec. 22 (2), 207–216.
URL https://doi.org/10.1145/170036.170072

Agrawal, R., Srikant, R., 1994. Fast algorithms for mining association rules in large databases. In: Proceedings of the 20th International Conference on Very Large Data Bases. VLDB '94. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 487–499.
URL https://dl.acm.org/citation.cfm?id=645920.672836

Bobadilla, J., Ortega, F., Hernando, A., Gutirrez, A., 2013. Recommender systems survey. Knowledge-Based Systems 46, 109 – 132.
URL https://doi.org/10.1016/j.knosys.2013.03.012

Boons, F., Chertow, M., Park, J., Spekkink, W., Shi, H., 2017. Industrial symbiosis dynamics and the problem of equivalence: Proposal for a comparative framework. Journal of Industrial Ecology 21 (4), 938–952.
URL https://doi.org/10.1111/jiec.12468

Burke, R., 2002. Hybrid recommender systems: Survey and experiments. User Modeling and User-Adapted Interaction 12 (4), 331–370.
URL https://doi.org/10.1023/A:1021240730564

Cecelja, F., Raafat, T., Trokanas, N., Innes, S., Smith, M., Yang, A., Zorgios, Y., Korkofygas, A., Kokossis, A., 2015. e-symbiosis: technology-enabled support for industrial symbiosis targeting small and medium enterprises and innovation. Journal of Cleaner Production 98, 336 – 352, special Volume: Support your future today! Turn environmental challenges into opportunities.
URL https://doi.org/10.1016/j.jclepro.2014.08.051

Chen, Y., Canny, J. F., 2011. Recommending ephemeral items at web scale. In: Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '11. ACM, New

York, NY, USA, pp. 1013–1022.
URL https://doi.org/10.1145/2009916.2010051

Chen, Z., Li, H., Kong, S. C., Hong, J., Xu, Q., 2006. E-commerce system simulation for construction and demolition waste exchange. Automation in Construction 15 (6), 706 – 718, knowledge Enabled Information System Applications in Construction.
URL https://doi.org/10.1016/j.autcon.2005.09.003

Chertow, M., Ehrenfeld, J., 2012. Organizing self-organizing systems. Journal of Industrial Ecology 16 (1), 13–27.
URL https://doi.org/10.1111/j.1530-9290.2011.00450.x

Chertow, M. R., 2000. Industrial symbiosis: literature and taxonomy. Annual review of energy and the environment 25 (1), 313–337.
URL https://doi.org/10.1146/annurev.energy.25.1.313

Chertow, M. R., 2007. "uncovering" industrial symbiosis. Journal of Industrial Ecology 11 (1), 11–30.
URL https://doi.org/10.1162/jiec.2007.1110

Clayton, A., Muirhead, J., Reichgelt, H., 2002. Enabling industrial symbiosis through a web-based waste exchange. Greener Management International 40, 93–107.

Cutaia, L., Luciano, A., Barberio, G., Sbaffoni, S., Mancuso, E., Scagliarino, C., La Monica, M., 2015. The experience of the first industrial symbiosis platform in italy. Environmental Engineering & Management Journal 14 (7), 1521–1533.

Dhanorkar, S., Donohue, K., Linderman, K., 2015. Repurposing materials and waste through online exchanges: overcoming the last hurdle. Production and Operations Management 24 (9), 1473–1493.
URL https://doi.org/10.1111/poms.12345

Dietrich, J., Becker, F., Nittka, T., Wabbels, M., Modoran, D., Kast, G., Williams, I., Curran, A., den Boer, E., Kopacek, B., et al., 2014. Extending product lifetimes: a reuse network for ict hardware. Proceedings of the ICE-Waste and Resource Management 167 (WR3), 123–135.
URL https://doi.org/10.1680/warm.13.00024

Ecoinvent, 2017. The ecoinvent database (version 3.3). https://www.ecoinvent.org/database/database.html, accessed: 2017-02-01.

Ekstrand, M. D., Riedl, J. T., Konstan, J. A., Feb. 2011. Collaborative filtering recommender systems. Found. Trends Hum.-Comput. Interact. 4 (2), 81–173.
URL https://doi.org/10.1561/1100000009

European Commission, 05 2000. Commission decision on the european list of waste (com 2000/532/ec). Tech. rep., European Commission.

European Commission, 2017. Horizon2020 project sharebox, unpublished workshop data.
URL https://www.sharebox-project.eu/

European Environmental Agency, 10 2016. More from less - material resource efficiency in europe. eea report no 10/2016. Tech. rep., European Environmental Agency.

Eurostat, 2018. Nace background. https://ec.europa.eu/eurostat/statistics-explained/ind accessed: 2018-03-22.

Freyne, J., Jacovi, M., Guy, I., Geyer, W., 2009. Increasing engagement through early recommender intervention. In: Proceedings of the Third ACM Conference on Recommender Systems. RecSys '09. ACM, New York, NY, USA, pp. 85–92.
URL https://doi.acm.org/10.1145/1639714.1639730

Geyer-Schulz, A., Hahsler, M., 2003. Comparing Two Recommender Algorithms with the Help of Recommendations by Peers. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 137–158.
URL httpss://doi.org/10.1007/978-3-540-39663-5_9

Gibbs, D., Deutz, P., 2007. Reflections on implementing industrial ecology through eco-industrial park development. Journal of Cleaner Production 15 (17), 1683 – 1695, from Material Flow Analysis to Material Flow Management.
URL https://doi.org/10.1016/j.jclepro.2007.02.003

Gomez-Uribe, C. A., Hunt, N., Dec. 2015. The netflix recommender system: Algorithms, business value, and innovation. ACM Trans. Manage. Inf. Syst. 6 (4), 13:1–13:19.
URL https://doi.org/10.1145/2843948

Grant, G. B., Seager, T. P., Massard, G., Nies, L., 2010. Information and communication technology for industrial symbiosis. Journal of Industrial Ecology 14 (5), 740–753.
URL https://doi.org/10.1111/j.1530-9290.2010.00273.x

Gregor, S., Hevner, A. R., Jun. 2013. Positioning and presenting design science research for maximum impact. MIS Q. 37 (2), 337–356.
URL https://doi.org/10.25300/MISQ/2013/37.2.01

Hazen, B. T., Boone, C. A., Ezell, J. D., Jones-Farmer, L. A., 2014. Data quality for data science, predictive analytics, and big data in supply chain management: An introduction to the problem and suggestions for research and applications. International Journal of Production Economics 154, 72 – 80.
URL https://doi.org/10.1016/j.ijpe.2014.04.018

Hein, A. M., Jankovic, M., Farel, R., Sam, L. I., Yannou, B., 2015. Modeling industrial symbiosis using design structure matrices. In: 17th international dependency and structure modeling conference, DSM 2015.

International Synergies Ltd., 2016. Training on industrial symbiosis taxonomies. Personal communication.

Jung, Y., Park, H., Du, D.-Z., Drake, B. L., 2003. A decision criterion for the optimal number of clusters in hierarchical clustering. Journal of Global Optimization 25 (1), 91–111.
URL https://doi.org/10.1023/A:1021394316112

Kimble, C., May 2013. Knowledge management, codification and tacit knowledge. Information Research 18 (12).
URL https://informationr.net/ir/18-2/paper577.html

Kogut, B., Zander, U., 1992. Knowledge of the firm, combinative capabilities, and the replication of technology. Organization Science 3 (3), 383–397.
URL https://doi.org/10.1287/orsc.3.3.383

Laybourn, P., Lombardi, D. R., 2012. Industrial symbiosis in european policy. Journal of Industrial Ecology 16 (1), 11.

Leong, Y. T., Tan, R. R., Aviso, K. B., Chew, I. M. L., 2016. Fuzzy analytic hierarchy process and targeting for inter-plant chilled and cooling water network synthesis. Journal of Cleaner Production 110, 40–53.
URL `https://doi.org/10.1016/j.jclepro.2015.02.036`

Lombardi, D. R., Laybourn, P., 2012. Redefining industrial symbiosis. Journal of Industrial Ecology 16 (1), 28–37.
URL `https://doi.org/10.1111/j.1530-9290.2011.00444.x`

Manning, C. D., Raghavan, P., Schütze, H., et al., 2008. Introduction to information retrieval. Vol. 1. Cambridge university press Cambridge.

Martnez-Ballesteros, M., Troncoso, A., Martnez-lvarez, F., Riquelme, J., 2016. Obtaining optimal quality measures for quantitative association rules. Neurocomputing 176 (Supplement C), 36 – 47, recent Advancements in Hybrid Artificial Intelligence Systems and its Application to Real-World Problems.
URL `httpss://doi.org/10.1016/j.neucom.2014.10.100`

Mirata, M., 2004. Experiences from early stages of a national industrial symbiosis programme in the uk: determinants and coordination challenges. Journal of Cleaner Production 12 (8-10), 967–983, applications of Industrial Ecology.
URL `https://doi.org/10.1016/j.jclepro.2004.02.031`

Natural Language Toolkit, 2017. Nltk 3.0 documentation. `https://www.nltk.org/`, accessed: 2017-01-31.

Natural Resources Wales, Scottish Environment Protection Agency, Northern Ireland Environment Agency, Environment Agency, 2015. Waste classification: Guidance on the classification and assessment of waste (1st edition 2015). Tech. rep.

Paquin, R. L., Howard-Grenville, J., 2012. The evolution of facilitated industrial symbiosis. Journal of Industrial Ecology 16 (1), 83–93.
URL `https://doi.org/10.1111/j.1530-9290.2011.00437.x`

Park, D. H., Kim, H. K., Choi, I. Y., Kim, J. K., 2012. A literature review and classification of recommender systems research. Expert Systems with Applications 39 (11), 10059 – 10072.
URL https://doi.org/10.1016/j.eswa.2012.02.038

Pathak, B., Garfinkel, R., Gopal, R., Venkatesan, R., Yin, F., Oct. 2010. Empirical analysis of the impact of recommender systems on sales. J. Manage. Inf. Syst. 27 (2), 159–188.
URL https://doi.org/10.2753/MIS0742-1222270205

Peffers, K., Tuunanen, T., Rothenberger, M. A., Chatterjee, S., 2007. A design science research methodology for information systems research. Journal of Management Information Systems 24 (3), 45–77.
URL https://doi.org/10.2753/MIS0742-1222240302

Poch, M., Comas, J., Rodrguez-Roda, I., Snchez-Marr, M., Corts, U., 2004. Designing and building real environmental decision support systems. Environmental Modelling & Software 19 (9), 857 – 873, environmental Sciences and Artificial Intelligence.
URL httpss://doi.org/10.1016/j.envsoft.2003.03.007

Porter, M. F., 1980. An algorithm for suffix stripping. Program 14 (3), 130–137.

Raafat, T., Trokanas, N., Cecelja, F., Bimi, X., 2013. An ontological approach towards enabling processing technologies participation in industrial symbiosis. Computers & Chemical Engineering 59, 33–46, selected papers from ESCAPE-22 (European Symposium on Computer Aided Process Engineering - 22), 17-20 June 2012, London, {UK}.
URL httpss://doi.org/10.1016/j.compchemeng.2013.03.022

Salvador, S., Chan, P., 2004. Determining the number of clusters/segments in hierarchical clustering/segmentation algorithms. In: Proceedings of the 16th IEEE International Conference on Tools with Artificial Intelligence. ICTAI '04. IEEE Computer Society, Washington, DC, USA, pp. 576–584.
URL https://doi.org/10.1109/ICTAI.2004.50

Sander, K., Schilling, S., Lskow, H., Gonser, J., Schwedtje, A., Kchen, V., 11 2008. Review of the european list of waste. Tech. rep., kopol GmbH and ARGUS GmbH.

Schoenherr, T., Griffith, D. A., Chandra, A., 2014. Knowledge management in supply chains: The role of explicit and tacit knowledge. Journal of Business Logistics 35 (2), 121–135.
URL https://doi.org/10.1111/jbl.12042

Shani, G., Gunawardana, A., 2011. Evaluating Recommendation Systems. Springer US, Boston, MA, pp. 257–297.
URL httpss://doi.org/10.1007/978-0-387-85820-3_8

Sheng, Y. P., Mykytyn Jr, P. P., Litecky, C. R., 2005. Competitor analysis and its defenses in the e-marketplace. Communications of the ACM 48 (8), 107–112.
URL https://doi.org/10.1016/j.jclepro.2015.02.036

Sterr, T., Ott, T., 2004. The industrial region as a promising unit for eco-industrial developmentreflections, practical experience and establishment of innovative instruments to support industrial ecology. Journal of Cleaner Production 12 (8-10), 947 – 965, applications of Industrial Ecology.
URL https://doi.org/10.1016/j.jclepro.2004.02.029

The ISDATA project, 2015. The industrial symbiosis data repository. https://isdata.org/, accessed: 2017-01-31.

United Nations Statistics Division, 2015. Central product classification (version 2.1). https://unstats.un.org/unsd/cr/registry/cpc-21.asp, accessed: 2017-01-31.

Van Beers, D., Corder, G., Bossilkov, A., Van Berkel, R., 2007. Industrial symbiosis in the australian minerals industry. Journal of Industrial Ecology 11 (1), 55–72.
URL https://doi.org/10.1016/j.mineng.2007.04.001

van Capelleveen, G., Amrit, C., Yazan, D. M., 2018. A Literature Survey of Information Systems Facilitating the Identification of Industrial Symbiosis. Springer International Publishing, Cham, pp. 155–169.
URL httpss://doi.org/10.1007/978-3-319-65687-8_14

Veiga, L. B. E., Magrini, A., 2009. Eco-industrial park development in rio de janeiro, brazil: a tool for sustainable development. Journal of Cleaner Production 17 (7), 653 – 661, present and Anticipated Demands for Natural Resources: Scientific, Technological, Political, Economic and Ethical

Approaches for Sustainable Management.
URL `https://doi.org/10.1016/j.jclepro.2008.11.009`

Wilcoxon, F., Wilcox, R. A., 1964. Some rapid approximate statistical procedures. Lederle Laboratories.

Wroblewska, A., Twardowski, B., Zawistowski, P., Ryżko, D., 2016. Automatic Clustering Methods of Offers in an E-Commerce Marketplace. Springer International Publishing, Cham, pp. 147–160.
URL `https://doi.org/10.1007/978-3-319-30315-4_13`

Yuan, X., Lee, J.-H., Kim, S.-J., Kim, Y.-H., 2013. Toward a user-oriented recommendation system for real estate websites. Information Systems 38 (2), 231 – 243.
URL `https://doi.org/10.1016/j.is.2012.08.004`

Zack, M. H., 1999. Managing codified knowledge. Sloan management review 40 (4), 45.

Appendix A. Appendix

Appendix A.1. Stem frequency vectorization

Algorithm 3: Stem-frequency vectorization of waste item-descriptions

---

    Data:

           N_{item} = Marketplace item to be vectorized

           W_{items} = Set of all waste items in marketplace

    Variables :

           U = Set of unique stems

    Result:

           V = Stem-frequency item vector

1 **Function** *Stem-Frequency-Vectorization(N_{item}, W_{items})*

2   Create a bag of stems N_{bag} for item N_{item};

3   Create a bag of stems W_{bags} for each item in W_{items};

4   **foreach** $bag \in W_{bags}$ **do**

5   **foreach** $stem \in bag$ **do**

6   **if** $stem \notin U$ **then**

7   Add the unique $stem$ to U;

8   Initialize an empty item vector V with n positions where n is equal to the number of unique stems U;

9   **foreach** stem $s \in N_{bag}$ **do**

10   Find the position p of stem s in the set of unique stems U;

11   Increase the frequency with $+1$ in the item vector V at position p ;

12   **return** V;

Algorithm 4: A simple multi-dimensional hierarchical agglomerative clustering algorithm

---

Data:

W_{items} = Set of all waste items in marketplace

Parameters:

max_it = Maximum number of iterations
min_sim = Minimum similarity cut-off criterium;

Variables   :

S = Matrix of cosine similarities
V = Stem-frequency item vector

Result:

C = List of clusters

1 **Function** *Simple-MHAC(W_{items}, max_it, min_sim)*
2   **foreach** *item* $\in W_{items}$ **do**
3   Create a stem-frequency vector V_{item} and store in V;
4   **for** $m \leftarrow 1$ **to** W_{items} **do**
5   **for** $n \leftarrow 1$ **to** W_{items} **do**
6   **if** $n > m$ **then**
7   Calculate the cosine similarity s for item combination
              $\{n,m\}$ using the associated item vectors V_n, and V_m and
              store the similarity $s_{\{m,n\}}$ in the similarity matrix S;

8   Assign item n to its own cluster $c_{\{n\}}$ in store cluster in C;
9   **for** *iteration* $\leftarrow 1$ **to** $W_{items} - 1$ **do**
10   Determine the key pair $\{m, n\}$ of the maximal similarity s_{max} in
         the matrix S.;
11   **if** *iteration* $<= max_it$ and $s_{max} >= min_sim$ **then**
12   Merge cluster C_m and C_n into a new cluster $C_{\{m,n\}}$;
13   Remove cluster C_m and C_n from C;
14   Remove similarities s_m and s_n from S;
15   Merge the item vector V_m and V_n into a new cluster vector
            $V_{\{m,n\}}$;
16   **foreach** *cluster* $\in C$ **do**
17   Calculate the cosine similarity s for item combination
               $\{n,m\}$ and *cluster* using the associated item vectors
               $V_{\{n,m\}}$, and $V_{cluster}$ and store the new similarity
               $s_{\{\{m,m\},cluster\}}$ in the similarity matrix S;

18   **return** C;