Indexing products of city centre retailers

Digitalizing products of city centre shops in order to improve online presence

Abstract
City centre retailers in many cases do not of have web-shops or a digitalized versions of their product catalogue. Consumer demands in retailing are changing and city centre retailers cannot match the new demands since their products are only visible in the stores. For this complication, a process design and algorithm is made which enables retailers to index their product in an efficient and reliable way by making using of EAN barcodes and web scraping techniques. The method showed that 82% of all products can be indexed correctly in 12.8 seconds average. The method and algorithm developed can be improved even more, this requires a larger sample and more advanced computational techniques.

1. Introduction
Since the introduction of the internet, a lot has been changed, including the retail sector. [1]: “The information economy has created more informed and demanding consumers than ever before. Successful retailers are responding to the needs of these customers by improving the trade-off between the customer benefits and transaction costs, thus creating superior customer value”. “Retailers started displaying and selling products through e-commerce channels. E-commerce definitely changed the retail business. Companies serve different channels and consumers behave different as well. Starting an online sales channel is profitable for a brick-and-mortar retailer [2,3]. Not all retailers are present online though, in particular the small and midsized retailers located in shopping streets or shopping malls [4] [5]. Those retailers see a decrease in revenue but encounter barriers in ad the online channel [6,7].

City centre retailers have none or limited (digital) information about their product catalogue [8]. Indexing their product catalogue is time consuming and thus expensive. In interviews retailers have indicated to only want to spend a limited amount of time on their online presence. On the other hand, consumers rate high quality images as the most important aspect of web shops. Furthermore, product name, price and specifications are rated as very important [8]. A clear dilemma is present: Consumer have high demands and retailers do not have the right information and do not want to invest much time on digitalizing this information. A suitable technology needed to be designed for this dilemma.

The goal of this study is to show well product information can be retrieved from the web by only using limited product information available at city centre stores. The objective of the algorithm is to index products from retailers in an efficient and reliable way. The demands for the technology are:

- Retailers must be able to index themselves
- Time to index a product should be less than one minute
- Product name must be accurate
- Product image must be of high quality

1.1 Research setup
This research consists out of six elements and is derived from general design science frameworks. To measure in which degree the goal of fast, reliable and efficient, several criteria have been drawn up. Those criteria are labour intensity, percentage auto indexed, percentage semi-automatic indexed and whether the data is available to index.

First a scan was done on the retailers of which sources of information were present. Secondly, the different information sources were analysed and the potential design space per source was examined. The most promising techniques were selected on which finally one algorithm have been developed further. Thirdly, an algorithm was developed and tested by gathering a random product samples from shops. The results for the different stores types and products categories are listed.

\[ \text{Goals and objectives} \]

\[ \text{criteria} \]

\[ \text{design space / Possible solution} \]

\[ \text{Testing designs on criteria} \]

\[ \text{Optimizing} \]

\[ \text{Optimizing} \]

Figure 1: research setup

2. Design space
First, a scan was done in order to assess which data is present which could serve as input for the digitization of products. The available information often only were the products itself, supplier lists. Some shops have a POS system with all possible products and some have actual product stock. A fast majority only had no actual product stock or only the products itself as information source.

Data sources on products are

1. The product itself(visual)
2. The tag attached to the product
   a. Name
   b. Barcode
   c. Specifications (optional)
   d. Some codes (optional)
3. A box (optional)
4. Tag placed by store owner (optional)

For the different sources product databases and various general search engines like Google were used to check which information was available. A various selection of search engines are used and only search engines which provided an API or other type of machine readable format since this is a requirement for further development steps.

The data source availability was assessed by determining how often that source is present at products in shops. The labour intensity of processing the data was also examined. This labour intensity is based on field research and the potential to automate. To measure labour time, timestamps were recorded on each entry. For example, barcodes processing could be automated very well. For each group, the results were listed, these numbers results the correct match percentage. All scores are rated on a scale from 1-5 where 5 is always the best score. A total overview of the scores is presented in Table 1.

As the Table 1 indicates, the barcodes have shown to be the best sources to build a product search engine on. The tag included more information which also found be quite accurate. Another promising solution is by using the store tags, those tags are generally easy accessible since they are placed on the product shelf while the barcode stickers on products were sometimes more time consuming to detect. However, not all shops have separate tags on shelves and in some cases, the barcode did not match the products barcode.

**Table 1: data/information scores**

<table>
<thead>
<tr>
<th>data source</th>
<th>Information source</th>
<th>availability</th>
<th>labour</th>
<th>result</th>
</tr>
</thead>
<tbody>
<tr>
<td>1 product visuals</td>
<td>Search engine</td>
<td>5</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>2 tag name</td>
<td>Search engine databases</td>
<td>4</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>3 Tag barcode</td>
<td>databases</td>
<td>5</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>4 Tag specs +name</td>
<td>Search engine</td>
<td>3</td>
<td>2</td>
<td>2</td>
</tr>
<tr>
<td>5 Tag codes</td>
<td>databases</td>
<td>4</td>
<td>4</td>
<td>1</td>
</tr>
<tr>
<td>7 Store tag</td>
<td>databases</td>
<td>3</td>
<td>5</td>
<td>3</td>
</tr>
</tbody>
</table>

When comparing the results, searching on barcodes is the most promising solution. There are several EAN code databases existing, paid and non-paid with or without an API. While conducting several tests with products, none of these database seems to provide enough information. When using the Google search website very good results were obtained. From most EAN codes, the correct name and picture were derived. This method proved to be successful, which led to the situation that no other methods were developed further.

2.1 Findings

For different scenario’s, other combination between data and information source can be used. From the product itself, a human description can be made which can be used as search query in the search box. It found to be very difficult to be precise in the description. The object name together with brand resulted in non-accurate results. Even if aspects like colour and material did not contribute to the accuracy.

Besides the goal of limiting efforts by retailers to index products, it also has a scientific and social goal. The scientific aspect is regarding information retrieval by searching, gathering storing and transforming data into information. Data on the web is growing exponential, therefore new techniques must be and can be developed to effectively make use of the exponential data growth. By developing techniques and algorithms to gather information, the efficiency of services can be improved.

The initial incentive for the use of barcodes was the growing use of more advanced point of sale systems which needed to rapidly scan barcodes in order to optimize retailing processes. With this experiment we will be able to show other positive effects of standardization; the reusability.

3. Sample gathering

For this sample two main aspects are needed. The barcode, and product information in order to validate the results. To do this, an image of the product is needed, together with the barcode. For this purpose a web-app is developed to index those products quickly. Using a Bluetooth barcode scanner, which functions as a keyboard is, paired with an android device. Using the latest HTML5 technologies by using the smartphone camera and storing images locally, images can be gathered with the corresponding barcode. On the image, the product names and looks must be present. The reason for storing the data locally is due to the data limits and upload speed. The image was stored as “product”.ean_shop_price.jpg. These images could later be processed and stored into a database.

In order to gather a representative sample, we used several randomizers in the process. By visiting various store types and store sizes the random element is introduced. The used randomizer were:

- From each product shelve or lane a total of two samples were taken.
- Product selected randomly on vertical and horizontal axis.

The grouping of the samples was done for 5 different product categories and price. Products types’ distinction in retailer literature is mainly based on two axis. First is the type of products and second...
is the price [9]. The samples gathering will be done by visiting several stores which primarily sell goods from this specific product type

- Shops in personal care
- Living and housing towards business
- Household products
- Shops in consumer electronics
- Do-it-yourself shops
- Education and leisure
- Retail non-in-shop

The process of sample gathering is schematically presented in Figure 2. Grey boxes represent human processes and white boxes represent automated backend processes. Web scraping from 10 different websites and then processing all image statistics takes 9.3 seconds on average. Therefore, this process is done asynchronously while scanning and not afterwards to spread the load on the servers. The duration of this processes is an important consideration when this technique would be applied on a larger scale.

Every single shop owner was willing to cooperate and some helped in the process. This willingness is important in order achieve the highest percentage of indexation. The gathering of samples from the product category “not-in-shop” was not conducted due to accessibility limitations. The accessibility of personal care products was also limited. Most jewellery is behind locked display cases. The retailer however was willing to open these cases and removing the price tags which were stickered over the barcodes. Stickers over barcodes were only seen at very small products with separate tags.

3.1 Processing
The first quick scan was done using the standard google search engine. Google provides an API for the search which returns a JavaScript object notation object which is easy to process using JavaScript. Retrieving accurate results using the google Custom Search API seemed promising but did not result in very accurate strings. While using the google API, the discovery has been made that the regular google search engine yield better results than the API’s. This discovery altered the process and from using the API, a scraping methodology was used which was far more effective.

The top 10 results often contained web page titles, names, advertisements or other random strings. For all individual words, the occurrences within the 10 results are counted. The assumption behind this is that words which have the most occurrences, are likely describing the product. The top 10 words are presented to a person. This person clicks on words in a specific order. In many cases, the order is automatically the correct. Also the images were shown where the correct image could be selected if it was present. If no proper images were found in the automated case, other methods were used. On average only 8 images were shown, this due to the fact that many websites have scraping protection or dynamic content.

Other aspects to further improve are language wise. In this case, the language was not defined which sometimes resulted in German results. However, the google algorithm knows the server origin, which was the Netherlands, which resulted in most keywords written in Dutch. In a few cases, results were listed from existing EAN databases.

The whole process of retrieving images based on product labels is shown in Figure 3.

4. Results
Three samples were not gathered correct. The barcode scanned did not result in a proper EAN or UPC code. In total 134 samples are taken where 132 were used for further processing. Not all samples gave good results, however, if a combination of different methods is used, the accuracy could be increased even more. This combination is between the full automated and semi-automatic processing.

4.1 Keywords
The results on the keyword search are shown in Table 2. The average score of keyword retrieval is 84%. Though the sample size per product type is small, the differences between them are not

![Figure 3: Sample processing process](image-url)
large. No statistical tests have been conducted since that is not the aim of the study and the sample size is quite limited.

In total 73% of the keywords could be found automatically based on the EAN codes, adding manual processing only yields an additional percentage of 11%. The needed time to index those products is 11 seconds versus 35 seconds. A trade-off between time consumption and added value can be made in a further design.

The results per price group are listed in Table 3. The total number of samples is recued to 116 since many products did not had a visible price tag. What can be concluded from the table is that more expensive products are easier to index. More in depth research could be done to explain this difference.

### Table 2: keyword retrieval results per product type

<table>
<thead>
<tr>
<th>Product Type</th>
<th>N</th>
<th>Keywords auto (%)</th>
<th>Keywords manual (%)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shops in personal care</td>
<td>5</td>
<td>80%</td>
<td>20%</td>
<td>100%</td>
</tr>
<tr>
<td>Living and housing towards business</td>
<td>16</td>
<td>81%</td>
<td>6%</td>
<td>88%</td>
</tr>
<tr>
<td>Household products</td>
<td>51</td>
<td>73%</td>
<td>6%</td>
<td>78%</td>
</tr>
<tr>
<td>Do-it-yourself shops</td>
<td>24</td>
<td>75%</td>
<td>17%</td>
<td>92%</td>
</tr>
<tr>
<td>Education and leisure</td>
<td>14</td>
<td>64%</td>
<td>7%</td>
<td>71%</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>73%</td>
<td>11%</td>
<td>84%</td>
</tr>
</tbody>
</table>

4.2 Images

Images are the most important element in online retailing and therefore included in this experiment. The problem with images is that is hard for software to determine the match of an image. It depends on background colour, size, objective display and correct version of product etc…For this simplicity purposes, no advanced image selection method was used, but purely by human eyes.

As Table 4 shows, the total of properly indexed products is 82%, the variation between the different product type groups is minimal. The sample from the education leisure products is biased since they are taken from a sewing shop and car parts. For these 2 groups, images are less important and is thus likely that less images of that product exist on the web. Also the names to describe those products are very general which does not lead to specific results. Since only images could be derived from products of which the system was able to identify keywords from, the scoring percentage of 82% is high. This means that only 2% of the products without a name, did not return a good image.

### Table 4: image indexing results per product type

<table>
<thead>
<tr>
<th>Product Type</th>
<th>N</th>
<th>Image auto (%)</th>
<th>Image (EAN) (%)</th>
<th>Image (name) (%)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Shops in personal care</td>
<td>5</td>
<td>40%</td>
<td>20%</td>
<td>40%</td>
<td>100%</td>
</tr>
<tr>
<td>Living and housing towards business</td>
<td>16</td>
<td>44%</td>
<td>25%</td>
<td>19%</td>
<td>88%</td>
</tr>
<tr>
<td>Household products</td>
<td>51</td>
<td>43%</td>
<td>20%</td>
<td>16%</td>
<td>78%</td>
</tr>
<tr>
<td>Do-it-yourself shops</td>
<td>24</td>
<td>50%</td>
<td>13%</td>
<td>25%</td>
<td>88%</td>
</tr>
<tr>
<td>Education and leisure</td>
<td>14</td>
<td>50%</td>
<td>7%</td>
<td>7%</td>
<td>64%</td>
</tr>
<tr>
<td>Total</td>
<td>134</td>
<td>44%</td>
<td>19%</td>
<td>19%</td>
<td>82%</td>
</tr>
</tbody>
</table>

The image retrieval methods show almost difference between product types and product price. If, and how strong, the correlation between product types is, is unknown. The selected product types are very broad defined and products within these groups can range as well. Example. A household product could be a washing machine or a bucket. As seen with the keywords, more expensive products can be indexed better.

### Table 5: Image indexing results per price group

<table>
<thead>
<tr>
<th>Price Range</th>
<th>N</th>
<th>Image (auto) (%)</th>
<th>Image (EAN) (%)</th>
<th>Image (name) (%)</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>0-5</td>
<td>24</td>
<td>38%</td>
<td>17%</td>
<td>21%</td>
<td>75%</td>
</tr>
<tr>
<td>5-10</td>
<td>18</td>
<td>39%</td>
<td>22%</td>
<td>17%</td>
<td>78%</td>
</tr>
<tr>
<td>10-20</td>
<td>16</td>
<td>38%</td>
<td>13%</td>
<td>25%</td>
<td>75%</td>
</tr>
<tr>
<td>20-50</td>
<td>15</td>
<td>53%</td>
<td>20%</td>
<td>20%</td>
<td>93%</td>
</tr>
<tr>
<td>50-100</td>
<td>6</td>
<td>50%</td>
<td>33%</td>
<td>17%</td>
<td>100%</td>
</tr>
<tr>
<td>100-200</td>
<td>1</td>
<td>100%</td>
<td>0%</td>
<td>0%</td>
<td>100%</td>
</tr>
<tr>
<td>200-500</td>
<td>28</td>
<td>46%</td>
<td>21%</td>
<td>18%</td>
<td>86%</td>
</tr>
<tr>
<td>500-1000</td>
<td>8</td>
<td>50%</td>
<td>0%</td>
<td>50%</td>
<td>100%</td>
</tr>
<tr>
<td>Total</td>
<td>116</td>
<td>44%</td>
<td>18%</td>
<td>22%</td>
<td>84%</td>
</tr>
</tbody>
</table>

While conducting the experiment, checks were done to compare the automatically found images versus images search on EAN codes versus images searched on name. In some cases searches on image names resulted the best picture while the automatically found images were also good enough. For the purposes of this experiment, the first good enough option was selected. When such an algorithm would be developed further, an automated check between sources could be made. The image search based on name, was found to be very accurate and thus promising for further research.

5. Conclusions

The goal of this research was to find a method which enables city centre retailers to easily index their products. That goal has been achieved since an average of 82% of all products can be correctly indexed with high quality images. The objective was to do this in an efficient and reliable way. The reliability was guaranteed with
the validating image and the comparison between methods. In reality, shop owners have to do this check, which is even more accurate since they know their product catalogue better. Regarding the efficiency, 73% of all products were indexed in an average of 11 seconds and in 11% of the cases, this is 35 seconds. This time represent the processing time, the sample gathering time was excluded. Though, considering an average product catalogue, indexing a shop takes between 0.5 and 3 days.

This experiment was conducted to estimate the match rate of automated product indexing and the time needed to index. The sample size is too small to draw hard conclusions on, especially between product groups, but it can be concluded that the method is suitable for city centre retailers of digitalize their store.

With a certain degree of certainty it can be said that more expensive products yield better results and that if keywords are found, images can be found.

6. Reflection & Recommendations

The used algorithm and process design is far from perfect, many improvements are possible. An improvement will not lead necessarily to higher success rates but to a time improvement both in product names and images. These improvements can be found in:

- Automatic correct product description without clicking.
- Take picture of product
  - Recognize barcode
  - Recognize other codes
- Take picture of multiple products

Web scraping is legal, though websites can detect scraping and disallow access, this detection was seen in the process as well. The automatically retrieval of images succeeded in an average of 8 websites. Scraping was also limited due to dynamic content loading. Google has a detection system for the google search pages. The practice of scraping the google search results is a common practice. Even Microsoft scraped google for enhancing its search engine Bing [10]. No official documentation is published about the google scraping abuse policies. From developers experiences the limit of scraping their website is 1000 requests per day per IP address. Various open source projects are present which are specifically designed to scrape the google results using IP-address spoofing and proxies to increase the search limit to near limitless [11].

Limitations of these research are mainly based on the sample size. A larger sample should be gathered in order to provide statistical prove and is necessary for further development of this method. Another limitation is that this algorithm relies on visible barcodes, but this is not the case in many product categories like food and second hand products.

7. References

8. Intomart GiK: Een andere kijk op online ondernemen. Deel II De online consument. (2013)

21. Yildiz, B., Miksch, S.: ontoX - A Method for Ontology-Driven Information Extraction. In Gervasi, O., Gavrilova, M.,