Exploiting Correspondence Analysis to Visualize Product Spaces

M. Giamattei, M. Scholz[1]

Abstract In this paper, correspondence analysis is proposed as visualization method for graphical product recommender systems and product catalogues. Since this method is applied to reduce a high-dimensional space containing products and attributes to a two-dimensional space, information loss is unavoidable. A simulation to measure the information loss is conducted in this article and shows that correspondence analysis is suitable to sustain at least 50 percent of the raw information when up to 8 attributes and an infinite number of products are visualized.

Introduction

Electronic shops offer online users vast collections of products. To overcome the amount of products and hence the plenty of product information in order to make a satisfying buying decision, several shops have implemented recommender systems [1]. These systems are suitable to narrow the set of available products, but they are not appropriate to point out which products and configurations of products are available. In contrast to recommender systems, product catalogues provide an overview of available products. Product catalogues mostly using a hierarchical order to present products, which prevents from identifying products having specific attribute levels. For example a notebook shop might categorize notebooks in a catalogue according to their screen size and the price. Identifying whether the shop provides notebooks with a specific processor speed, price and memory size is not possible. Applying two-dimensional coordinate systems allows displaying products described by only two attributes. Several studies concerning the use of attributes to make a buying decision have shown that consumers consider between 3 and 8 attributes in average. Increasing the number of attributes in a coordinate system is thus worthwhile. Though, adding further visual dimensions to a coordinate system is not practicable. Therefore, a method for dimension reduction is required which can for example map 8 dimensions (one per attribute) to a two-dimensional display.

In this paper, we propose correspondence analysis as method for visualizing multiple product attributes, their relations (such as conflicts and concords) and the

[1] University of Passau, Junior Professorship of Information Systems, Passau, Germany, marcus@giamattei.de, michael.scholz@uni-passau.de

assignment of products to the attributes. Correspondence analysis has already been used for information visualization and is hence not a novel approach. However, correspondence analysis has never been used for visualizing attributes and products in a product catalogue or recommender system. The first research question we investigate in this paper is hence the following:

RQ1: How to use correspondence analysis for visualizing product attributes and products in a two dimensional space?

Since we use correspondence analysis to visualize an *n*-dimensional space in a two-dimensional display, an information loss is not avoidable. If for instance only 20 percent of the raw information content is tenable when reducing an 8-dimensional space to a two-dimensional space, correspondence analysis is inappropriate as visualization method in product recommender systems or product catalogues. This lets us formulate the next research question for this paper:

RQ2: How well do correspondence analysis adhere information?

To answer the above mentioned research questions the remainder is organized as follows. In the next section we review on related work and discuss how our method can complement existing ones. Afterwards, we briefly introduce correspondence analysis and show how to interpret the output of such an analysis for a product recommender system or a product catalogue. We then estimate the information loss of the correspondence analysis depending on the number of visualized attributes. We conclude our paper with limitations and implications for future research.

Related Work

In this section, we briefly review existing methods and systems providing a visualization of product or attribute relationships to either navigate through a product space or to get product recommendations. According to the items the methods visualize, we can distinguish between methods visualizing products and their relations, attributes and their relations, and products as well as attributes and their relations. Correspondence analysis is suitable to visualize both products and attributes with their relations in one space.

Visualizing products and their relations requires estimating similarities between products. In statistics, multidimensional scaling (MDS) is known "as a method that represents (dis)similarity data as distances in a low-dimensional space in order to make these data accessible to visual inspection and exploration" [2]. Kagie et al. developed a visualization system based on MDS which enables users to ex-

plore a product space and to find similar products for a given one [3]. Since this system only visualizes products with their relations, a user does not know why two products are similar and furthermore a user can not utilize the system to get an overview of possible combinations of attribute levels (product profiles). In contrast to MDS, a correspondence analysis can be used to plot two variables (products and attributes) in one view.

Another method for displaying multi-dimensional data sets is a parallel coordinate system [4]. A visual interface for product catalogues based on parallel coordinate systems has been proposed in [5]. This system enables users to easily explore which attribute levels are available for a particular product category and which attribute levels a specific product is consisting of. Although all attribute levels are visible, it is hard to recognize which attributes are opposite and which are compatible.

A visualization based on a two-dimensional coordinate systems has been presented by Ahlberg and Shneiderman [6]. The system is suitable to visualize three attributes (one per coordinate axis and one with different colors). Relationships between products are apparent according to the attributes used in the visualization. In their prototype, they visualize movies with the attributes years, popularity and genre. Further attributes as well as attribute relationships are not representable. In contrast to the system proposed in [5], the prototype of Ahlberg and Shneiderman does not allow its users to formulate preferences which are imperative for product recommendation.

Presenting products and attributes in a two-dimensional view is conceivable using a table. This method allows users to easily compare products based on their attribute levels. Though, a table is not practicable for visualizing relationships between products or attributes. Spenke et al. presented an interactive system for product comparison and selection based on a table [7].

Theetranont et al. have investigated in a system which shows a three-dimensional product space [8]. By viewing the product space generated by the system, users can gain an understanding of the relationships between product attributes. Since the system is based on multi-attribute utility theory, it is also applicable for product recommendation.

The idea of visualizing a product space to get an overview of existing products or to better select products is not novel and as shown in this section several work has already be done (for an overview see [9]). In this paper, we investigate in correspondence analysis (CA) as visualization method because:

- no recent work has been incorporate CA as a method to visualize product spaces,
- in contrast to other methods (e.g. MDS or parallel coordinates), a CA is suitable to present both products and attributes with their relationships in one view,
- several improvements for CA exist (e.g. to visualize more than two variables in one view)

Correspondence Analysis

Correspondence analysis is a statistical method that is used to explore associations between the variables of a cross-tabulation. It is a geometric method for visualizing variables and their associations in a low-dimensional space. Mapping on to a low-dimensional space allows exploring the data structure, which is similar in nature to principal component analysis (PCA). In contrast to principal component analysis, CA can also be used for analyzing the structure of categorical data. The name is a translation of the French "Analyse des Correspondances", which means analysis of a system of associations. It was developed by Benzérci in the 1960's and 1970's [10], but it just gained widespread popularity after Greenacre published the method in English [11].

A CA begins with a cross-tabulation containing two variables (e.g. products and product attributes). The computation proceeds in three steps: data standardization, dimension extraction and normalization [12]. Table 1 shows a cross-tabulation for notebooks. For each attribute and product, the absolute number of consumers who assigned the attribute to the product is given.

Table 1. Example of a cross-tabulation

Product	Performance	Mobility	Handling	Built-Quality
P1	5	7	7	4
P2	4	5	8	9
P3	7	3	9	6
P4	8	2	5	5

The core of a CA is the dimension extraction. In this step two dimensions are extracted from the cross-tabulation. For each level of the two variables (product attributes and products), their value for each dimension is also extracted. This step is mostly conducted using singular value decomposition. The output of the dimension extraction is afterwards normalized. Each attribute level and each product is now described by a value for each of the two extracted dimensions. Plotting the attributes and products in a two-dimensional coordinate system is thus possible. The plot of the output of a CA of Table 1 is shown in Figure 1.

Figure 1 shows that product P1 is very close to mobility and product P4 is close to performance. It also shows that mobility and performance have a large distance and are hence conflicting attributes while performance nearly concurs with built-quality. A consumer can see that it is not possible to get a notebook with high mobility and high performance, while products with good handling and good built-quality exist (P2 and P3). Such visualizations have the following advantages for users of product recommender systems or product catalogues:

- it supports users by identifying which product profiles are available on an electronic shop,
- it supports users by reliably building preferences for product attributes (since the users can identify attribute trade-offs),
- it supports users by comparing two or more products based on their attribute levels.

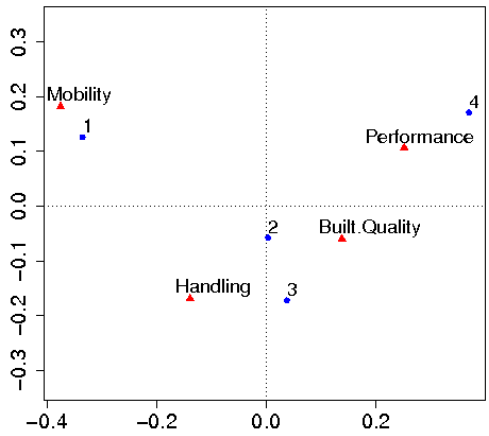

Fig. 1. Two-dimensional plot of the correspondence analysis of Table 1

Such an interpretation is, however, only valid if most of the information of the raw data is to be kept in the plot of a correspondence analysis. In the next section, we examine the information loss with a simulation.

Information Loss of Correspondence Analysis

The information loss created by the CA can be measured very easily. The singular value decomposition selects the two dimensions which cover the highest amount of information. Other dimensions are cancelled. Information here is equivalent to variance and can be measured best by the inertia. The inertia has the advantage to be independent from the number of cases (here the number of persons having rated the notebooks). The inertia T is defined as follows, where r is the number of ratings conducted to produce the cross-tabulation and χ^2 the statistical standard measure for dispersion:

$$T = \frac{\chi^2}{r}$$

The percentage of the Eigenvalues PEW_k has to be computed for each dimension k to assess the information covered by a CA. This can be done by dividing the

singular value s_k of the dimension k through the overall inertia T. The information content γ of the plot is the sum of the values for the first and second dimension.

$$PEW_k = \frac{s_k}{T}$$
$$\gamma = PEW_1 + PEW_2$$

Taking this into account, the information loss can be depicted as $1\text{-}\gamma$. We use content of information γ instead of the information loss $1\text{-}\gamma$. A simulation was conducted to test the information content for different settings. The number of products was constant at 20 products because in the CA the minor number of attributes or products determines the information loss. The number of attributes varied from 3 to 20. These numbers were chosen because the information loss starts with more than three dimensions and a reasonable product search will not work with more than 15 or 20 attributes. For every possible level of attributes 10,000 matrices were created at random. Every matrix contained pseudo-random numbers in the range from 0 to 9. Other settings showed that this range does not affect the information loss but decreases the inertia[2]. Therefore, this description is focused only on the range mentioned first. For every set the content of information γ was assessed as described before.

For a reasonable classification of the results the maximal information loss has to be computed in order to compare it with the experimental data. The maximal information loss is bounded upwards, because the intertia is it, too. The maximal intertia is the minimum of the numbers of attributes and products. Thus, the minimal content of information γ_{min} can be computed as follows:

$$\gamma_{min} = \frac{2}{\min(products, attributes) - 1}$$

This is due to the fact, that the maximal inertia has to be distributed equally over the dimensions. Taking for example five dimensions, the content of information is 20 % per dimension and 40 % for the first and the second dimension.

As shown in Table 2 still 50 % of the raw information adheres in the plot when using 10 attributes. 57 % of the information is to be kept when displaying up to 8 attributes. Several studies about the consumer information behavior indicate that not more than 8 attributes and normally only 6 attributes are considered for a buying decision by a consumer [13, 14, 15]. The number of considered attributes has been reported for different products, such as toothpaste, coffee, cameras and other electronic equipment. Restricting the number of attributes according to the consumers' cognitive boundaries is hence valid.

The information content does not fit a normal distribution for any amount of attributes (proved with a Kolmogrov-Smirnov-test). To get a picture of the distri-

[2] Tested for the ranges [0..49] and [0..99].

bution, we computed the skewness υ_1 and the kurtosis β_2 for each amount of attributes. As shown in Table 2 the information content is a value of a distribution with a kurtosis equal to a normal distribution ($\beta_2 \approx 3$) but a positive skew. Thus, in most cases the information content is lower than the expected mean. For example, when 8 attributes are visualized, we can ascertain that in 5555 of 10000 cases the information content was lower than the estimated mean of 0.57804. However, in more than 8900 cases the information content is higher than 0.5 when 8 attributes are displayed. We thus can adhere to the rule that up to 8 attributes are presentable with less than 50 % of information loss.

Table 2. Mean, standard deviation, skewness, kurtosis and minimal information of a CA

Attributes	γ	SD(γ)	$\upsilon 1(\gamma)$	$\beta 2(\gamma)$	γ_{min}
3	1.00000	0.00000	-	-	1.00000
4	0.86150	0.06753	0.23069	2.29200	0.66667
5	0.76108	0.08057	0.48339	2.72950	0.50000
6	0.68434	0.07798	0.58073	2.98896	0.40000
7	0.62658	0.07383	0.58927	3.05911	0.33333
8	0.57804	0.06720	0.58317	3.09736	0.28571
9	0.53918	0.06043	0.53606	3.05092	0.25000
10	0.50630	0.05428	0.48333	3.03195	0.22222
11	0.47791	0.04885	0.44049	3.01068	0.20000
12	0.45288	0.04385	0.43086	3.03176	0.18182
13	0.43128	0.03851	0.35757	2.98788	0.16667
14	0.41160	0.03497	0.36426	3.09175	0.15385
15	0.39411	0.03176	0.33855	3.15533	0.14286
16	0.37782	0.02883	0.29247	3.10903	0.13333
17	0.36453	0.02679	0.31497	3.02763	0.12500
18	0.35106	0.02428	0.30383	3.11559	0.11765
19	0.33917	0.02264	0.28867	3.04673	0.11111
20	0.32811	0.02145	0.30182	3.17788	0.10526

Conclusion

In this paper, we have shown how correspondence analysis could be used in a product recommender system or product catalogue. We also conducted a simulation in order to proof whether enough information adhere when visualizing a variable number of attributes. According to our findings, we can state that correspondence analysis is a promising method for visualizing product spaces. However, an experimental investigation of a concrete implementation is missing and will be hence conducted in future.

The results of our simulation indicate that the information loss is less than 50 % when conducting a correspondence analysis based on 8 attributes and *n* products. The products used in the simulation are hypothetical products consisting of randomly generated attribute levels. Real products are often more close to each other than the hypothetical products. We therefore assume that we have overestimated the information loss.

For practitioners, our findings indicate that a visualization based on correspondence analysis would be an appropriate extension for an electronic shop. The major advantage for consumers is the increased support in finding adequate products.

As shown in section 2, other methods have been investigated for graphical shopping systems. Comparing our approach with the existing systems will be also necessary in future research. We suggest comparing these systems by their predictive quality, the user satisfaction and the information loss.

References

1. Schafer, J.B., Konstan, J.A., Riedl, J. (2001) E-Commerce Recommendation Applications, *Journal of Data Mining and Knowledge Discovery* 5(1): 115-153.
2. Borg, I. and Groenen, P.J. (2005) *Modern Multidimensional Scaling – Theory and Application*, 2nd edition. New York, Springer.
3. Kagie, M., van Wezel, M. and Groenen, P.J. (2008) A Graphical Shopping Interface Based on Product Attributes, *Decision Support Systems* 46(1): 265-276.
4. Klemz, B.R. and Dunne, P.M. (2000) Exploratory Analysis Using Parallel Coordinate Systems – Data Visualization in n-Dimensions, *Marketing Letters* 11(4): 323-333.
5. Lee, J., Lee, H.S. and Wang, P. (2004) An Interactive Visual Interface for Online Product Catalogs, *Electronic Commerce Research* 4: 335-358.
6. Ahlberg, C. and Shneiderman, B. (1994) Visual Information Seeking – Tight Coupling of Dynamic Query Filters with Starfield Displays, *ACM SIGCHI Conference on Human Factors in Computing Systems*.
7. Spenke, M., Beilken, C. and Berlage, T. (1996) Focus – The Interactive Table for Product Comparison and Selection, *ACM Symposium on User Interface Software and Technology*.
8. Theetranont, C., Haddawy, P. and Krairit, D. (2007) Integrating Visualization and Multi-Attribute Utility Theory for Online Product Selection, *International Journal of Information Technology & Decision Making* 6(4): 723-750.
9. Hearst, M. (2009) *Search User Interfaces*. Cambridge, Cambridge University Press.
10. Benzérci, J.P. (1963) *Course de Linguistique Mathématique*. Rennes, Universitée de Rennes.
11. Greenacre, M. (1984) *Theory and Applications of Correspondence Analysis*. London, Academic Publishing.
12. Greenacre, M. (2007) *Correspondence Analysis in Practice,* 2nd edition. Boca Raton, Chapman & Hall.
13. Sheluga, D.A., Jaccard, J. and Jacoby, J. (1979) Preference, Search, and Choice – An Integrative Approach, *Journal of Consumer Research* 6(2): 166-176.
14. Ratchford, B.T. and van Raaij, W. (1980) Information Acquisition Process and Monetary Loss due to Incorrect Choice, 5th Annual Colloquium on Economic Psychology.
15. Kroeber-Riel, W. and Weinberg, P. (2003) *Konsumentenverhalten*, 8th edition. Munich, Vahlen.