Choosing Attribute Weights for Item Dissimilarity using Clikstream Data with an Application to a Product Catalog Map

Martijn Kagie, Michiel van Wezel and Patrick J.F. Groenen

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2008-024-MKT
Publication	April 2008
Number of pages	14
Persistent paper URL	http://hdl.handle.net/1765/12243
Email address corresponding author	kagie@few.eur.nl
Address	Erasmus Research Institute of Management (ERIM) RSM Erasmus University / Erasmus School of Economics Erasmus Universiteit Rotterdam P.O.Box 1738 3000 DR Rotterdam, The Netherlands Phone: + 31 10 408 1182 Fax: + 31 10 408 9640 Email: info@erim.eur.nl Internet: www.erim.eur.nl

Bibliographic data and classifications of all the ERIM reports are also available on the ERIM website:
www.erim.eur.nl

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES
RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS

Abstract	In content- and knowledge-based recommender systems often a measure of (dis)similarity between items is used. Frequently, this measure is based on the attributes of the items. However, which attributes are important for the users of the system remains an important question to answer. In this paper, we present an approach to determine attribute weights in a dissimilarity measure using clickstream data of an e-commerce website. Counted is how many times products are sold and based on this a Poisson regression model is estimated. Estimates of this model are then used to determine the attribute weights in the dissimilarity measure. We show an application of this approach on a product catalog of MP3 players provided by Compare Group, owner of the Dutch price comparison site http://www.vergelijk.nl, and show how the dissimilarity measure can be used to improve 2D product catalog visualizations.
Free Keywords	clickstream data, comparison, attribute weights, dissimilarity measure
Availability	The ERIM Report Series is distributed through the following platforms: Academic Repository at Erasmus University (DEAR), DEAR ERIM Series Portal Social Science Research Network (SSRN), SSRN ERIM Series Webpage Research Papers in Economics (REPEC), REPEC ERIM Series Webpage
Classifications	The electronic versions of the papers in the ERIM report Series contain bibliographic metadata by the following classification systems: Library of Congress Classification, (LCC) LCC Webpage Journal of Economic Literature, (JEL), JEL Webpage ACM Computing Classification System CCS Webpage Inspec Classification scheme (ICS), ICS Webpage

Choosing Attribute Weights for Item Dissimilarity using Clikstream Data with an Application to a Product Catalog Map

Martijn Kagie Michiel van Wezel Patrick J.F. Groenen
Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
The Netherlands
{kagie,mvanwezel,groenen}@few.eur.nl

April 24, 2008

Abstract

In content- and knowledge-based recommender systems often a measure of (dis)similarity between items is used. Frequently, this measure is based on the attributes of the items. However, which attributes are important for the users of the system remains an important question to answer. In this paper, we present an approach to determine attribute weights in a dissimilarity measure using clickstream data of an e-commerce website. Counted is how many times products are sold and based on this a Poisson regression model is estimated. Estimates of this model are then used to determine the attribute weights in the dissimilarity measure. We show an application of this approach on a product catalog of MP3 players provided by Compare Group, owner of the Dutch price comparison site http://www.vergelijk.nl, and show how the dissimilarity measure can be used to improve 2D product catalog visualizations.

1 Introduction

Many content- or knowledge-based recommender systems [3] use some type of case-based reasoning or nearest neighbor retrieval [14, 16]. These techniques heavily rely on some (dis)similarity measure between different items for their recommendation strategy. Often, this similarity measure is based on the attributes of the items. However, not all attributes of an item are equally important to the user and, thus, in the recommendation process. Therefore, the measure of similarity should use some type of attribute weighting. Otherwise, the similarity measure used in the system will not match the notion of similarity between items the users have and, thus, the system will recommend the wrong items.

Although weights are generally specified by experts, some work has been done on recommender systems that automatically learn these weights user specifically. Schwab et al. [21] learn user specific weights for binary features using significance testing assuming normal distributions. When the user selects items having a specific attribute value more or less often, that is, there is a significant effect, this attribute got a higher or lower weight. Arslan et al. [1] used the number of times an attribute was used in the query of the user to learn these attribute weights. Finally, Coyle and Cunningham [6] compare the final choice of the user with the provided recommendations and learn the feature weights from that.

All these approaches assume that the user gives the system time to let it learn his/her preferences in one or more sessions. However, in many e-commerce domains users expect immediate appropriate recommendations and for a large group of product categories, such as durable goods, users will not come back to buy such a product within a couple of years.

In this paper, we introduce a generic way to choose attribute weights. We use a dissimilarity measure that can handle both different kinds of attributes and missing values. The attribute weights are estimated using clickstream logs of an e-commerce site. In these log files, we count how often each item was sold. Based on the assumption that attributes that have a high influence on the sales of products are attributes that are considered to be important by the user, we estimate a Poisson regression model [17, 15] on sales and product attributes.

Besides in recommender systems, dissimilarity between products has also been used in map based e-commerce interfaces [12, 11]. In our paper, we discuss an improved prototype of the interface introduced in [12] which uses the weighted dissimilarity measure. This prototype and the weighted dissimilarity measure are applied to a product catalog of MP3 players. Both product data and clickstream files were provided by Compare Group, owner of the Dutch price comparison site `http://www.vergelijk.nl`.

The remainder of the paper is organized as follows. In the next section, we introduce the dissimilarity measure for which the weights are determined. In Section 3, we describe the Poisson regression model, how we handle missing values in this model, and how the results of the Poisson regression model are used to create weights for the dissimilarity measure. Then, in Section 4, we show an application of the attribute weight determination on a product catalog of MP3 players and show how this methodology can be applied in a map based user interface. Finally, in Section 5, we draw conclusions and indicate directions for future research.

2 Dissimilarity

First we introduce the measure we use to compute dissimilarity between products. To this end, we introduce some notation. Consider a data set D, which contains n products having K attributes $\{(x_{i1}, x_{i2} \ldots, x_{iK})\}_1^n$. For each product, we also have a binary vector $\mathbf{m}_i = (m_{i1}, m_{i2} \ldots, m_{iK})$, containing values of 1 for nonmissing attribute values. In most applications, these attributes have mixed types, that is, the attributes can be numerical, binary, or categorical.

The most often used (dis)similarity measures, like the Euclidean distance, Pearson's correlation coefficient, and Jaccard's similarity measure, are only suited to handle one of these attribute types. Also, these measures cannot cope with missing values in a natural way. Therefore, we use a dissimilarity measure which is based on the general coefficient of similarity proposed by Gower [7], which was also used in Kagie et al. [11].

The dissimilarity δ_{ij} between products i and j is defined as the square root of the weighted average of nonmissing dissimilarity scores δ_{ijk} on the K attributes

$$\delta_{ij} = \sqrt{\frac{\sum_{k=1}^{K} w_k m_{ik} m_{jk} \delta_{ijk}}{\sum_{k=1}^{K} w_k m_{ik} m_{jk}}} \ , \tag{1}$$

in which the w_k's are the weights for the different dissimilarity scores and, hence, for the different attributes. These weights w_k specify how important the different attributes are in the computation of the dissimilarity measure and, hence, in the application. In the next section, we will discuss how we determine these weights based on an approach using clickstream data.

The computation of the dissimilarity scores δ_{ijk} in (1) is dependent on the type of the attribute. For numerical attributes, the dissimilarity score δ_{ijk} is the normalized absolute distance

$$\delta_{ijk}^{N} = \frac{|x_{ik} - x_{jk}|}{\left(\sum_{i<j} m_{ik} m_{jk}\right)^{-1} \sum_{i<j} m_{ik} m_{jk} |x_{ik} - x_{jk}|} \ . \tag{2}$$

For categorical attributes, the dissimilarity score δ_{ijk} is defined as

$$\delta_{ijk}^{C} = \frac{1(x_{ik} \neq x_{jk})}{\left(\sum_{i<j} m_{ik} m_{jk}\right)^{-1} \sum_{i<j} m_{ik} m_{jk} 1(x_{ik} \neq x_{jk})} \ , \tag{3}$$

where 1() is the indicator function returning a value of 1 when the condition is true and 0 otherwise.

However, in many product catalogs, as will also be the case in the catalog used in this paper, a third type of attributes exist, which we call multicategorical attributes. Where a product can have only one value for a categorical attribute such as, for example, its brand, it can have multiple values for a multicategorical attribute. For instance, an MP3 player can have an attribute called ‘supported audio formats’, which can contain the values MP3 and WMA at the same time.

We assume that two products are identical on a multicategorical attribute, when they share exactly the same values. So, we propose to compute the dissimilarity score for a multicategorical attribute by counting the number of values that only one of the products has. More formally, we can define the dissimilarity score δ_{ijk} for multicategorical attributes as

$$\delta_{ijk}^{M} = \frac{|x_{ik} \cup x_{jk}| - |x_{ik} \cap x_{jk}|}{\left(\sum_{i<j} m_{ik} m_{jk}\right)^{-1} \sum_{i<j} m_{ik} m_{jk} (|x_{ik} \cup x_{jk}| - |x_{ik} \cap x_{jk}|)} \ , \tag{4}$$

where both x_{ik} and x_{jk} are sets of values. Note that this leads to identical results as when we represent every unique attribute value by a binary variable and then count the unequal values for two products, that is, computing the Hamming distance between these binary variables. However, using (4) the total number of unique values is not needed to compute the dissimilarity score.

3 Choosing Attribute Weights

In the previous section, we introduced the dissimilarity measure for which we like to determine the weights w_k for the different attributes. In this section, we will introduce an approach to determine these weights using clickstream data. For every product, we count how often it was sold during some period. Using these counts and the product attributes, we estimate a Poisson regression model, which is a model belonging to the class of generalized linear models. Using the coefficients of this model and their corresponding standard errors, we compute t-values which form the basis of our attribute weights.

A very popular group of models in the field of statistics are the generalized linear models (GLM) [17, 15]. Most well-known models belonging to this class are the linear regression and logistic regression model. Again, we have our data set D, having items $\{\mathbf{x}_i\}_i^n$. These items still have K attributes. In GLMs we cannot use (multi)categorical attributes directly, so we have to create dummy variables instead. Therefore, every categorical attribute is represented by L_k dummies $x_{ik\ell}$, which are 1 for the category where the item belongs to and 0 for all other attributes, where L_k is the number of different categories for attribute k minus one (this is done to avoid multicollinearity). When an item belongs to the last category ($L_k + 1$) all dummies for this attribute will be 0. For multicategorical attributes the same approach is used, only now all categories are represented by the L_k dummies. For, numerical attributes we have only one variable that represents the attribute. Hence, $x_{ik} = x_{ik1}$ and $L_k = 1$. We collect all $x_{ik\ell}$ for item i in vector $\mathbf{x}_i$. Also, an intercept term x_{i0} is incorporated in this vector, which equals 1 for all items. Furthermore, we have an independent variable value y_i for all n items. Now, we can express the group of GLMs as

$$y_i \approx f(\mathbf{x}_i'\mathbf{b}) \ , \tag{5}$$

where $f()$ is some function and $\mathbf{b}$ is a vector of regression parameters.

Different GLMs can be made by specifying different functions $f()$ and assuming different distributions from the exponential family for y_i having expectation $E(f(\mathbf{x}_i'\mathbf{b}))$ in (5). For instance, specifying $f(\theta) = \theta$ and assuming a normal distribution leads to the ordinary linear regression model. In our application, dependent variable y will contain counts of sales for different products. Since y in that case is discrete and nonnegative, the specification of ordinary linear regression will be incorrect. Therefore, we will use another type of model from the GLM family, namely the Poisson regression model, which is often used for count data. The Poisson regression model is specified by

$$y_i \approx \exp(\mathbf{x}_i'\mathbf{b}) \ , \tag{6}$$

where we assume that y_i has a Poisson distribution. Note that in the Poisson regression model $f(\theta) = \exp(\theta)$. All GLMs can be trained by maximizing their corresponding loglikelihood function. Often, this is done by an iteratively reweighted least squares algorithm.

One serious drawback of the Poisson regression model (and other GLMs) for our application is that it lacks an integrated way of handling missing values, while product catalogs often contain a lot of missing values, since producers all supply different attributes. Imbrahim et al. [10] recently compared different techniques that can be used to handle missing values in combination with GLMs.

One of the best methods (leading to unbiased estimates and reliable standard errors) in their comparison was multiple imputation (MI) [19]. MI methods create Q 'complete' data sets in which values for originally missing values are drawn from a distribution conditionally on the nonmissing values. Methods to create these imputed data sets are data augmentation [20] and sampling importance/resampling [13]. Both methods lead to imputations of the same quality, where the second method needs substantially less computation time. Therefore, we will use the second method, more specifically the Amelia algorithm [13] which is available as a package [9] for statistical software environment `R`.

When using Q imputed data sets, the GLM, in our case the Poisson regression model, has to be estimated on all Q data sets. Following [19] we can compute estimates of regression coefficients and standard errors. The estimate for a regression coefficient $b_{k\ell}$ then becomes

$$b_{k\ell} = \frac{1}{Q} \sum_{q=1}^{Q} b_{k\ell q} \ , \tag{7}$$

where $b_{k\ell q}$ is the estimate of $b_{k\ell}$ on the q-th imputed data set. Note that this is just the average for $b_{k\ell}$ over the Q imputed data sets. Computation of standard errors is less straightforward, since these should include both the uncertainty in the specific GLMs and the uncertainty introduced by the imputations. Therefore, the estimated standard error $\sigma_{k\ell}$, more specifically the estimated variance $\sigma^2_{k\ell}$, consists of a part measuring the within-imputation variance $SW_{k\ell}$ and a part measuring between-imputation variance $SB_{k\ell}$. The within-imputation variance is computed in the following way

$$SW_{k\ell} = \frac{1}{Q} \sum_{q=1}^{Q} \sigma^2_{k\ell q} \ , \tag{8}$$

where $\sigma^2_{k\ell q}$ is the estimated variance of $b_{k\ell}$ on the q-th data set, which follows from the Poisson regression procedure. The between-imputation variance is specified as follows

$$SB_{k\ell} = \frac{1}{Q-1} \sum_{q=1}^{Q} (b_{k\ell q} - b_{k\ell})^2 \ . \tag{9}$$

Finally, following [19], both parts are combined to compute the total estimated variance of $b_{k\ell}$

$$\sigma^2_{k\ell} = SW_{k\ell} + \left(1 + \frac{1}{Q}\right) SB_{k\ell} \ . \tag{10}$$

The estimated standard errors $\sigma_{k\ell} = \sqrt{\sigma^2_{k\ell}}$, can be used to compute t-values in the usual way

$$t_{k\ell} = \frac{b_{k\ell}}{\sigma_{k\ell}} \ . \tag{11}$$

The resulting coefficients $b_{k\ell}$ from the Poisson regression model cannot be used directly as weights in the dissimilarity measure (1) for several reasons. The first reason is that the scales of the dissimilarity scores and variables are not the same. Second, when using $b_{k\ell}$ directly as weight for the corresponding dissimilarity score, we do not take into account the uncertainty we have about the correctness of this coefficient. Although a coefficient can be relatively high, it can still be unimportant. For example, this can be the case with dummies having very few 1's. Then, this high impact of the coefficient is only applicable to a limited number of items and its total importance is limited. By taking the uncertainty we have into account, we can correct for this. Finally, we want to have $w_k \geq 0$, while $b_{k\ell}$ can also be negative, when a certain variable has a negative influence on the sales of a product.

The first two problems that exist when using $b_{k\ell}$ as weight in the dissimilarity measure can be overcome by using the t-value $t_{k\ell}$ of coefficient $b_{k\ell}$ as basis of the weight computation. Since the $t_{k\ell}$'s are standardized they are comparable to each other as are the dissimilarity scores. Since this standardization is done by division of the corresponding standard error $\sigma_{k\ell}$, uncertainty about $b_{k\ell}$ is incorporated into $t_{k\ell}$. When we use $|t_{k\ell}|$ instead of $t_{k\ell}$ we guarantee the weights to be larger than or equal to 0. This can be done, because it does not matter for the importance of an attribute in the dissimilarity whether the influence of the attribute is positive or negative, but the size of this influence does.

When attribute k is numerical, we can map $|t_{k1}|$ (i.e. $\ell = 1$) directly to the a 'pseudo' absolute t-value v_k for attribute k, that is, $v_k = |t_{k1}|$. Then, including a normalization of the weights (for ease of interpretability), we can compute w_k using

$$w_k = \frac{v_k}{\sum_{k'=1}^{K} v_{k'}} \quad . \tag{12}$$

For (multi)categorical attributes, we first have to compute v_k using the L_k values of $t_{k\ell}$. This is done by taking the average of the absolute $t_{k\ell}$ values

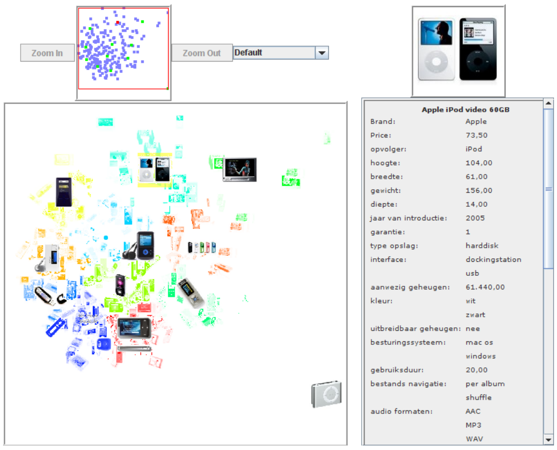
$$v_k = \frac{1}{L_k} \sum_{\ell=1}^{L_k} |t_{k\ell}| \quad . \tag{13}$$

These v_k's can then be used to compute the weights for (multi)categorical attributes using (12).

The t-values $t_{k\ell}$ can be compared to a t-distribution having

$$df_{k\ell} = (Q-1)\left(1 + \frac{Q \cdot SW_{k\ell}}{(Q+1)SB_{k\ell}}\right)^2 \tag{14}$$

degrees of freedom to determine p-values for hypothesis testing. These p-values can be used in a so-called stepwise model. A stepwise model performs variable selection only keeping the variables that have a statistically significant effect on y, that is, having a $b_{k\ell}$ statistically different from 0. To definitely find the 'best' model one would have to compare models with all different combinations of variables. In practice, this is often computationally infeasible. Therefore, stepwise approaches take a greedy approach by starting with a model containing all variables and then, each time, deleting the most insignificant variable. Note that this is not the same as immediately deleting all insignificant variables, since due to collinearity significance of variables may change when deleting another variable from the model. When using the stepwise model to determine weights w_k, we consider L_k to be the number of dummies incorporated in the final model. Since it is not clear whether using a stepwise model leading to less attributes having all relatively higher weights in the dissimilarity measure will



Apple iPod video 60GB

Brand:	Apple
Price:	73,50
opvolger:	iPod
hoogte:	104,00
breedte:	61,00
gewicht:	156,00
diepte:	14,00
jaar van introductie:	2005
garantie:	1
type opslag:	harddisk
interface:	dockingstation usb
aanwezig geheugen:	61.440,00
kleur:	wit zwart
uitbreidbaar geheugen:	nee
besturingssysteem:	mac os windows
gebruiksduur:	20,00
bestands navigatie:	per album shuffle
audio formaten:	AAC MP3 WAV

Figure 1: GUI of the 2D Product Map interface.

lead to better results than a model containing all variables, we consider both models in our evaluation.

4 Application to MP3 Players

Now we have introduced the techniques we use to create the weighted dissimilarity measures, we show an application of these dissimilarities on a product catalog of MP3 players. Also, we show the implications of using these weighted dissimilarities in creating a product map using the methodology used by Kagie et al. [12]. They introduced an online shopping interface based on a 2D map of the product catalog that is made using a technique called multidimensional scaling (MDS) [2]. MDS creates these maps based on a matrix of dissimilarities. A screenshot of this GUI is show in Figure 1.

In the product map a limited set of products is highlighted by giving them a larger full color image. Which products are highlighted is determined by a k-means clustering as described in [12]. The user can explore the map by zooming in and out on different parts of the map. Furthermore, the user can label the products by attribute values or popularity additionally to the default labeling by cluster.

Both the product catalog and the clickstream log files were made available to us by Compare Group. Compare Group hosts, among other European price comparison sites, the Dutch price comparison site `http://www.vergelijk.nl`. The product catalog used is based on a data base dump of this site from October 2007. The log files are used to count how many times users clicked on a link to an internet shop to buy a certain product, which is called an ‘outclick’. We counted these ‘outclicks’ during two months from July 15 until September 15, 2007. Since the product catalog changes over time, the data set used to determine the attribute weights is slightly different than the product catalog used in the prototype. For the determination of the weights a data set is used containing all MP3 players that were sold (‘outclicked’) at least one time during the two months analyzed and could be matched to product attributes available

Table 1: Attribute characteristics of the data used to estimate the attribute weights. For (multi)categorical attributes only the three most occurring values are shown.

Attribute	% Missing	Mean
y	0.0%	109.18
Numerical Attributes		
Price	4.4%	141.06
Height	40.8%	63.96
Width	40.8%	47.66
Weight	44.7%	69.71
Depth	40.8%	17.26
Memory Size	0.4%	10315.56
Battery Life	44.7%	18.69
Categorical Attributes		
Brand	0.0%	Philips (12.3%), Samsung (11.4%), Creative (8.3%)
Radio	34.7%	yes (69.8%), no (30.2%)
Extendable Memory	47.4%	yes (13.3%), no (86.7%)
Equalizer	39.0%	yes (85.6%), no (14.4%)
Screen	29.4%	yes (99.4%), no (0.6%)
Battery Type	45.2%	li-ion accu (40.8%), 1×AAA (36.0%), li-polymer (20.8%)
Voice Memo	23.3%	yes (81.7%), no (18.3%)
Multicategorical Attributes		
Memory Type	42.5%	flash memory (68.7%), harddisk (21.4%), sd card (9.2%)
Interface	4.0%	usb (65.3%), usb 2.0 (29.7%), hi-speed usb (6.9%)
Color	38.6%	black (65.7%), white (20.0%), silver (17.9%)
Operating System	30.7%	windows (79.8%), mac os (34.2%), windows xp (29.8%)
Audio Formats	1.8%	MP3 (98.7%), WMA (90.2%), WAV (47.8%)

in the database (the data base contained except products that are sold now, also old products). This lead to a data set of 228 MP3 players that is summarized in Table 1. Although the original database contained more product attributes than there are shown in the table, these attributes were not used in the analysis, since they have more than 50% missing values. This is done, since estimation of the missing values of these attributes becomes very hard and, since they are hardly observed, these attributes most likely do not have a significant impact on the sales of a product. Furthermore, to make the imputation of variables easier we excluded the dummy variables of categories that were observed less than 10 times.

We estimated the parameters of Poisson regression models using the statistical software environment `R` [18]. First, we created 25 imputed data sets using the Amelia package [9]. Then, using the built-in `R` function `glm`, 25 Poisson regression models were estimated on the imputed data sets. Although 3–5 imputations are considered to be enough in many applications [19], we use 25, since the data has a very high degree of missingness. For the stepwise model this process was repeated as described in Section 3.

The estimated model coefficients of the stepwise Poisson regression model are shown in Table 2. The deletion of variables was stopped when all remaining coefficients had a p-value of 0.05 or lower. Besides the coefficient estimates $b_{k\ell}$, the table also shows the corresponding standard error $\sigma_{k\ell}$ and t-value $t_{k\ell}$. Finally, it shows the v_k and corresponding weight w_k for all attributes represented in the model.

As can be seen in the table, the brand of the product is the most important attribute identifying popularity of a product in the MP3 market. A-brand MP3 players are sold up to 54 times than more than MP3 players of regular brands *ceteris paribus*. (For binary (dummy) variables y is $\exp(b_{k\ell})$ times larger when this variable is 1 rather than 0, other things being equal. [22]) Also, memory size

Table 2: Coefficients of the stepwise Poisson regression model.

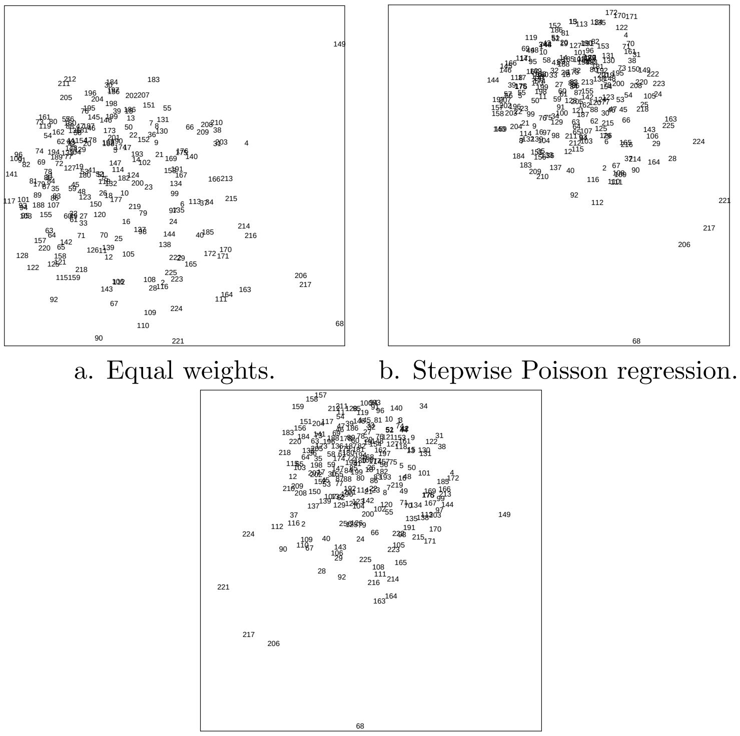
Attribute	Category	$b_{k\ell}$	$\sigma_{k\ell}$	$t_{k\ell}$	v_k	w_k
Intercept		2.24	0.15	14.77		
Brand					8.76	0.242
	Apple	3.97	0.20	19.77		
	Creative	2.53	0.23	11.22		
	Philips	0.79	0.14	5.78		
	Samsung	1.15	0.20	5.71		
	Sandisk	1.43	0.21	6.94		
	iAudio	0.71	0.23	3.13		
Width		-0.01	0.00	-3.25	3.25	0.090
Memory Type					3.05	0.084
	HDD	0.86	0.28	3.05		
Memory Size		-0.00	0.00	-6.38	6.38	0.176
Color					3.06	0.084
	white	0.50	0.18	2.81		
	black	0.53	0.16	3.30		
OS					3.11	0.086
	windows vista	-1.44	0.46	-3.11		
Audio-form.					4.75	0.131
	atrac3	1.05	0.22	4.75		
Battery					3.83	0.106
	li-ion-accu	0.74	0.19	3.83		

Table 3: Attribute weights based on full Poisson regression model.

	v_k	w_k
Brand	2.36	0.119
Price	0.80	0.040
Height	1.78	0.090
Width	0.96	0.048
Weight	0.18	0.009
Depth	1.61	0.081
Radio	0.33	0.017
Memory Type	0.81	0.041
Interface	0.30	0.015
Memory Size	2.50	0.125
Color	1.13	0.057
Extendable Memory	1.03	0.052
Operating System	0.97	0.049
Battery Life	0.71	0.036
Audio Formats	0.74	0.037
Equalizer	0.50	0.025
Screen	0.42	0.021
Battery Type	0.89	0.045
Voice Memo	1.85	0.093

has a high impact. Customers seem to prefer MP3 players with smaller amounts of memory. There are two coefficients that need some more explanation: *OS: Windows Vista* and *Audio Format: Atrac3*. Both effects seem somewhat odd at first sight. However, the negative effect of Windows Vista support may be caused by the fact that MP3 Players supporting Windows Vista are relative new models and were maybe not available during the complete two month period. The Atrac3 audio format was introduced by Sony and is poorly adopted by other brands. Although this effect of Atrac3 is stronger than the influence of the Sony brand, it is possible that this is indeed an effect belonging to *Brand* and not to *Audio Format*.

As mentioned earlier, we also estimated a full Poisson regression model. The v_k's and weights w_k estimated using this model are shown in Table 3. Also, using this model the attributes *Brand* and *Memory Size* are considered the most important attributes getting the highest weights. However, due to the fact that there are more variables considered in the model, the absolute weights of these



a. Equal weights.

b. Stepwise Poisson regression.

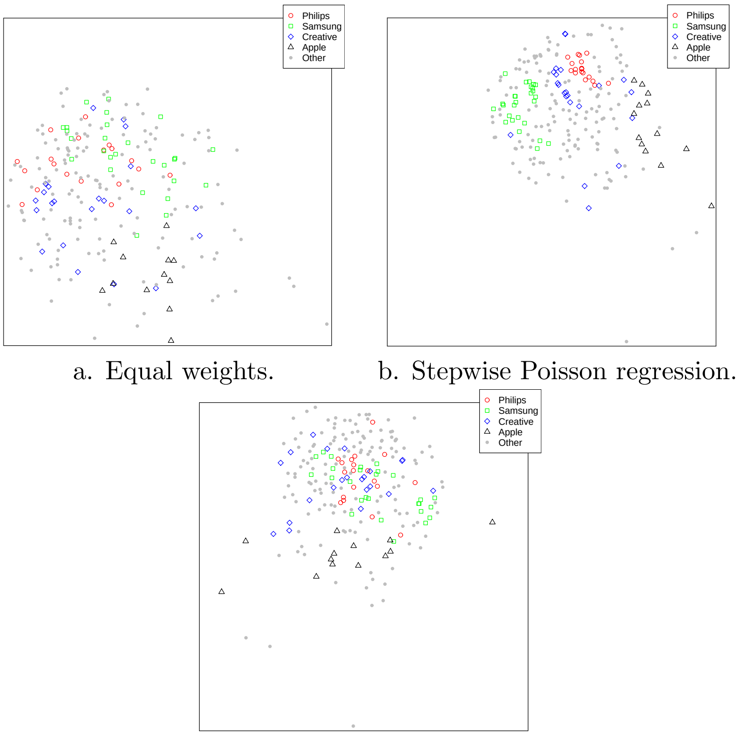
c. Full Poisson regression.

Figure 2: Product maps using the different weighting schemes. Points are labeled by the case numbers of the products.

attributes are lower than in the stepwise Poisson regression model.

Figure 2a shows the product map created using the original approach that was described in [12]. All attributes are considered equally important (all weights are set to 1) and all attributes are used in the computation of the dissimilarity measure, also the attributes that have so many missing values that they were excluded from the Poisson regression analysis. All points are labeled by the case number of the corresponding product. The product maps created using the stepwise and full Poisson regression model are shown in Figures 2b and 2c. These maps are rotated using Procrustean transformations [8] to best map the original unweighted map. To get more insight in these three different maps, we advise to try the prototypes implementing these three weighting schemes that are available on `http://people.few.eur.nl/kagie/wprodmaps.htm`.

However, we also provide somewhat more insight in these maps here. Figures 3a–3c show the three previously showed product maps only labeled by their brands. In both the stepwise and full Poisson regression model *Brand* was the attribute getting the highest weight and this should have an influence on the resulting maps. As can be seen in Figure 3b, the use of the stepwise Poisson regression weights leads to a map in which the products belonging to a single brand are almost all clustered together. The clustering on brand in Figure 3c is less strong, as may be expected since the brand weight was lower, but also in this map the clustering is stronger than in the original unweighted map. Interesting



a. Equal weights.

b. Stepwise Poisson regression.

c. Full Poisson regression.

Figure 3: Product maps labeled by brand for the different weighting schemes.

to see is that contrary to the single clusters of brands in the stepwise map, products of the same brand that are relative similar are now clustered together, leading to more clusters for one brand on different places in the map. For instance, when we have a look at the Creative MP3 players, we see that there are is a different cluster for the Creative Zen Vision models at the bottom left which are quite large and have a large memory size, while the smaller models such as the Zen V and Nano models are clustered in the middle of the map. Since the effect of the important attributes on the map seems stronger in the stepwise approach,it seems that this method should be preferred, although user tests should be conducted to be more certain.

5 Conclusions and Discussion

In this paper, we introduced a generic way to estimate attribute weights for dissimilarity computation for e-commerce product catalogs using clickstream data. In the clickstream logs for each product was counted how often it was sold. Then, a Poisson regression model was used to estimate how much influence the different attributes have on the sales of the products. Using the coefficients of this model, attribute weights for the dissimilarity were computed. We compared two Poisson regression models. One model containing only significant coefficients and a full model containing all attributes.

Both models indicated the brand and the memory size of a product as best indicators for its popularity. These effects were stronger in the stepwise model (with only significant attributes) than in the full Poisson regression model, because insignificant attributes correlating with these attributes where excluded from this model.

The weights resulting from both models were used to create product maps of a product catalog of MP3 players to be used in a map based shopping interface as introduced in [12]. Both the stepwise and the full Poisson regression approach lead to maps in which products were more clustered based on the important attributes as was expected. This effect was stronger using the weights that resulted from the stepwise model.

An important line for future research is to compare the new weighting approach with the unweighted approach in real user experiments. Not only, we intend to do this for the map based interface, but also in a recommender system context.

Furthermore, this approach could be used on slightly different kind of data using different models from the GLM family. When rating data is available linear regression could be used and when there are also negative examples (not liked products) binary logistic regression might be an option. Also, these models can be extended using latent classes to provide user specific estimates.

A drawback of this type of linear models is that interaction effects are not incorporated in these models. Therefore, the resulting weights might be biased. A line for future research therefore might be to use models that can model interaction effects, such as generalized regression trees [5, 4].

Acknowledgements

We thank Compare Group for making their product catalog and clickstream log files available to us.

References

[1] B. Arslan, F. Ricci, N. Mirzadeh, and A. Venturini. A dynamic approach to feature weighting. *Management Information Systems*, 6:999–1008, 2002.

[2] I. Borg and P. J. F. Groenen. *Modern Multidimensional Scaling.* Springer Series in Statistics. Springer, New York, 2nd edition, 2005.

[3] R. Burke. Knowledge based recommender systems. In J. E. Daily, A. Kent, and H. Lancour, editors, *Encyclopedia of Library and Information Science*, volume 69, Supplement 32. Marcel Dekker, New York, 2000.

[4] P. Chaudhuri, W.-D. Lo, W.-Y. Loh, and C.-C. Yang. Generalized regression trees. *Statistica Sinica*, 5:641–666, 1995.

[5] A. Ciampi. Generalized regression trees. *Computational Statistics & Data Analysis*, 12:57–78, 1991.

[6] L. Coyle and P. Cunningham. Improving recommendation rankings by learning personal feature weights. In *Advances in Case-Based Reasoning; 7th European Conference, ECCBR 2004. Proceedings.*, volume 3155 of *Lecture Notes in Computer Science*, pages 560–572, Springer, Heidelberg, 2004.

[7] J. C. Gower. A general coefficient of similarity and some of its properties. *Biometrics*, 27:857 – 874, 1971.

[8] B. F. Green. The orthogonal approximation of an oblique structure in factor analysis. *Psychometrika*, 17:429–440, 1952.

[9] J. Honaker, G. King, and M. Blackwell. *Amelia II: A Program for Missing Data*, 2008. R package version 1.1-27, http://gking.harvard.edu/amelia.

[10] J. G. Ibrahim, M.-H. Chen, S. R. Lipsitz, and A. H. Herring. Missing-data methods for generalized linear models: A comparative review. *Journal of the American Statistical Association*, 100(469):332–346, 2005.

[11] M. Kagie, M. van Wezel, and P. J. F. Groenen. A graphical shopping interface based on product characteristics. In V. Oria, A. Elmagarmid, F. Lochovsky, and Y. Saygin, editors, *Proceedings of the 23rd International Conference on Data Engineering Workshops*, pages 791–800. IEEE Computer Society, 2007.

[12] M. Kagie, M. van Wezel, and P. J. F. Groenen. Online shopping using a two dimensional product map. In G. Psaila and R. Wagner, editors, *E-Commerce and Web Technologies; 8th International Conference, EC-Web 2007. Proceedings.*, volume 4655 of *Lecture Notes in Computer Science*, pages 89–98. Springer, Heidelberg, 2007.

[13] G. King, J. Honaker, A. Joseph, and K. Scheve. Analyzing incomplete political science data: An alternative algorithm for multiple imputation. *American Political Science Review*, 95(1):49–69, 2001.

[14] F. Lorenzi and F. Ricci. Case-based recommender systems: A unifying view. In B. Mobasher and S. S. Anand, editors, *Intelligent Techniques for Web Personalization*, volume 3169 of *Lecture Notes in Computer Science*, pages 89–113. Springer, Heidelberg, 2005.

[15] P. McCullagh and J. A. Nelder. *Generalized Linear Models*, volume 37 of *Monographs on Statistics and Applied Probability.* Chapman & Hall, Boca Raton, 2nd edition, 1989.

[16] D. McSherry. A generalised approach to similarity-based retrieval in recommender systems. *Artificial Intelligence Review*, 18:309–341, 2002.

[17] J. A. Nelder and R. W. M. Wedderburn. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[18] R Development Core Team. *R: A language and environment for statistical computing.* R Foundation for Statistical Computing, Vienna, Austria, 2008. `http://www.R-project.org`.

[19] D. B. Rubin. *Multiple Imputation for Nonresponse in Surveys.* Wiley, New York, 1987.

[20] J. L. Schafer and M. K. Olsen. Multiple imputation for multivariate missing-data problems: A data analyst's perspective. *Multivariate Behavioral Research*, 33(4):545–571, 1998.

[21] I. Schwab, W. Pohl, and I. Koychev. Learning to recommend from positive evidence. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pages 241–246. ACM Press, 2000.

[22] M. Verbeek. *A Guide To Modern Econometrics.* John Wiley & Sons, Chichester, UK, 2nd edition, 2004.

Publications in the Report Series Research[*] in Management

ERIM Research Program: “Marketing”

2008

Experts' Stated Behavior
Youssef Boulaksil and Philip Hans Franses
ERS-2008-001-MKT
http://hdl.handle.net/1765/10900

The Value of Analogical Reasoning for the Design of Creative Sales Promotion Campaigns: A Case-Based Reasoning Approach
Niek A.P. Althuizen and Berend Wierenga
ERS-2008-006-MKT
http://hdl.handle.net/1765/11289

Shopping Context and Consumers' Mental Representation of Complex Shopping Trip Decision Problems
Benedict G.C. Dellaert, Theo A. Arentze and Harry J.P. Timmermans
ERS-2008-016-MKT
http://hdl.handle.net/1765/11812

Modeling the Effectiveness of Hourly Direct-Response Radio Commercials
Meltem Kiygi Calli, Marcel Weverbergh and Philip Hans Franses
ERS-2008-019-MKT
http://hdl.handle.net/1765/12242

Choosing Attribute Weights for Item Dissimilarity using Clikstream Data with an Application to a Product Catalog Map
Martijn Kagie, Michiel van Wezel and Patrick J.F. Groenen
ERS-2008-024-MKT
http://hdl.handle.net/1765/12243

[*] A complete overview of the ERIM Report Series Research in Management:
https://ep.eur.nl/handle/1765/1

ERIM Research Programs:

LIS	Business Processes, Logistics and Information Systems
ORG	Organizing for Performance
MKT	Marketing
F&A	Finance and Accounting
STR	Strategy and Entrepreneurship