An Empirical Comparison of Dissimilarity Measures for Recommender Systems

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

ERIM REPORT SERIES *RESEARCH IN MANAGEMENT*	
ERIM Report Series reference number	ERS-2009-023-MKT
Publication	May 2009
Number of pages	16
Persistent paper URL	http://hdl.handle.net/1765/15911
Email address corresponding author	kagie@ese.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

ERASMUS RESEARCH INSTITUTE OF MANAGEMENT

REPORT SERIES
RESEARCH IN MANAGEMENT

ABSTRACT AND KEYWORDS

Abstract	Many content-based recommendation approaches are based on a dissimilarity measure based on the product attributes. In this paper, we evaluate four dissimilarity measures for product recommendation using an online survey. In this survey, we asked users to specify which products they considered to be relevant recommendations given a reference product. We used microwave ovens as product category. Based on these responses, we create a relative relevance matrix we use to evaluate the dissimilarity measures with. Also, we use this matrix to estimate weights to be used in the dissimilarity measures. In this way, we evaluate four dissimilarity measures: the Euclidean Distance, the Hamming Distance, the Heterogeneous Euclidean-Overlap Metric, and the Adapted Gower Coefficient. The evaluation shows that these weights improve recommendation performance. Furthermore, the experiments indicate that when recommending a single product, the Heterogeneous Euclidean-Overlap Metric should be used and when recommending more than one product the Adapted Gower Coefficient is the best alternative. Finally, we compare these dissimilarity measures with a collaborative method and show that this method performs worse than the dissimilarity based approaches.
Free Keywords	dissimilarity, case-based recommendation, evaluation, weight estimation
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

An Empirical Comparison of Dissimilarity Measures for Recommender Systems

Martijn Kagie* Michiel van Wezel Patrick J.F. Groenen
Econometric Institute
Erasmus School of Economics
Erasmus University Rotterdam
The Netherlands

May 7, 2009

Abstract

Many content-based recommendation approaches are based on a dissimilarity measure based on the product attributes. In this paper, we evaluate four dissimilarity measures for product recommendation using an online survey. In this survey, we asked users to specify which products they considered to be relevant recommendations given a reference product. We used microwave ovens as product category. Based on these responses, we create a relative relevance matrix we use to evaluate the dissimilarity measures with. Also, we use this matrix to estimate weights to be used in the dissimilarity measures. In this way, we evaluate four dissimilarity measures: the Euclidean Distance, the Hamming Distance, the Heterogeneous Euclidean-Overlap Metric, and the Adapted Gower Coefficient.

The evaluation shows that these weights improve recommendation performance. Furthermore, the experiments indicate that when recommending a single product, the Heterogeneous Euclidean-Overlap Metric should be used and when recommending more than one product the Adapted Gower Coefficient is the best alternative. Finally, we compare these dissimilarity measures with a collaborative method and show that this method performs worse than the dissimilarity based approaches.

Key words: Dissimilarity, Case-Based Recommendation, Evaluation, Weight Estimation.

1 Introduction

A wide variety of approaches has been proposed during the last decade to recommend products from a product catalog to users. These approaches are known as recommender systems [24]. In general, three different types of recommender systems are distinguished [2]: Collaborative, content-based, and hybrid methods. Collaborative filtering recommends products based on similarity of the user's taste with the taste of other users. In contrast, content-based methods use

*Corresponding author. E-mail: kagie@ese.eur.nl characteristics of products to base their recommendations on. Hybrid methods combine both approaches.

Although collaborative methods are more popular, content-based methods are far more useful in certain electronic commerce domains, such as consumer electronics and other durable goods. In these domains, not only products can be described by a well defined set of product characteristics, but there is also only limited user preference data available, since people buy these products infrequently.

On the basis of a large group of these content-based methods, the case-based methods [19, 20] lies some measure of dissimilarity between products and/or products and queries. In this paper, we will evaluate a number of state-of-the-art case-based dissimilarity concepts using a novel evaluation method based on an online survey. In this survey, people were asked to select relevant recommendations given a reference product. Based on the responses, we could determine which products users considered to be the most relevant recommendations given a reference product. We define a measure, called mean average relative relevance (MARR), to evaluate recommendation lists produced by some dissimilarity measure, which measures the relevance of the recommendations provided by this dissimilarity measure. Also, we use the relevance of recommendations to estimate weights in the dissimilarity measure to improve performance of the dissimilarity measures.

The remainder of the paper is organized as follows. In the next section, we discuss some work related to this paper. Then, in Section 3 we introduce the dissimilarity measures evaluated in this paper and, in Section 3.5, we discuss the way we determine optimal weights for these measures. In Section 4, we introduce the survey setup and the MARR measure and apply these to evaluate the dissimilarity measures on a product catalog of microwave ovens in Section 5. Finally, we draw conclusions and give directions for future research.

2 Related Work

As mentioned in the introduction, there exist three types of recommender systems [2]: Collaborative filtering [10], content-based recommendation [22], and hybrid recommendation [8], combining these two approaches.

Collaborative filtering is the most popular recommendation method both in science and industry (for example, Amazon.com uses collaborative filtering to provide recommendations [18]). In collaborative filtering, recommendation is based on similarity in taste between users. To measure taste, this kind of systems collect ratings of products by users. These ratings can be explicitly been given by the users or implicit ratings based on, for example, surfing behavior. Many collaborative filtering approaches, the so-called memory based methods [2, 6], use a similarity measure between users [23] or products [27] to provide recommendations. Hence, no product characteristics are used.

In contrast, content-based methods (and knowledge-based methods [7]) use the characteristics of the products in the computation of their recommendations. In fact, content-based methods recommend products that are similar to the products a users liked in the past. How this is actually done, depends heavily on the kind of data available. When there is a quite large user history available containing both possitive (liked) as negative (disliked) examples and the products have structured attributes, almost every machine learning algorithm could be used to provide the recommendations [22], such as decision trees [16], nearest neighbor [4], and naive Bayes [21].

However, often only positive examples are available, such as viewed products or bought products. In such circumstances, the use of machine learning techniques is more problematic and products most similar to products seen or bought in the past are recommended. For this purpose, a (dis)similarity measure is needed. This approach is known as case-based recommendation [20], since these systems are based on methodology from case-based reasoning [1]. Also, in case-based reasoning similarity plays a key role and a lot of different (dis)similarity measures where proposed and used in that field (for an overview, see [17]). Of these, Euclidean and Hamming distance are the most widely used [17] and, therefore, these measures are incorporated in our study as benchmark measures.

Commercial product catalogs demand some specific properties from the dissimilarity measure used. Not only are products in these databases specified by attributes of different types (numerical, categorical), but often there is also a lot of missing data. Two dissimilarity measures that both deal with these aspects and were recently proposed in the recommender system literature are considered in this paper. The Heterogeneous Euclidean-Overlap Metric [30] that has been used in recommender system applications in [3, 25, 26] can handle numerical and categorical attributes and missing values. The Adapted Gower Coefficient [14] can also handle these, but also supports multi-valued categorical attributes.

Most of the time, users do not consider all attributes of a product to be equally important while making a purchase decision. For instance, in the case of MP3 players brand is an important characteristic users generally base their decisions on, while the signal-to-noise ratio is unimportant for most users. To overcome this, weights can be incorporated in the dissimilarity measure. Although weights are generally specified by experts, some work has been done on recommender systems that automatically learn these weights user specifically, such as [3, 5, 9, 28]. Often, these user specific approaches are based on weighting approaches originating from case-based reasoning [29].

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 domains such as consumer electronics this is not the case. In [13] and [15], we proposed two methods to derive weights using product popularity, for example, measured as sales of a product. Although evaluations [15] showed that these kinds of weighting may improve recommendation performance, it was not clear which method performed best. The methods have the advantage that they only rely on some measure of product popularity to determine attribute weights and are able to handle missing values. In this paper, we introduce another approach, which uses the recommendation relevance matrix to estimate these weights as is discussed in Section 3.5.

Although, as stated earlier, a lot of dissimilarity approaches exist in case-based reasoning and recommender systems literature, we are unaware of an evaluation comparing these dissimilarity measures on real commercial data sets. Herlocker et al.[12] evaluated different kinds of collaborative filtering algorithms on the well-known MovieLens data set. However, since they did not consider product characteristics, also content-based approaches were not evaluated. Zanker et al. [31], on the other hand, compared some knowledge-based, collaborative filtering, and one content-based method on a commercial catalog containing cigars. However, this evaluation was entirely based on clickstream analysis, which has the disadvantage that it is unclear how to distinguish between user preferences and the website structure.

3 Dissimilarity Measures

In this section, we discuss the dissimilarity measures that are evaluated in this paper. To this end, we introduce the following notation. Consider a product catalog, which contains n products having K attributes $\{(x_{i1}, x_{i2}, \ldots, x_{iK})\}_1^n$. In the electronic commerce domain, these attributes may have mixed types, that is, the attributes can be, for example, numerical, binary, or categorical. Furthermore, we introduce weights w_k for the different attributes, such that attributes that are considered important can be given a higher weight. Finally, we will use δ_{ij} as dissimilarity between items i and j respectively. In some cases, the dissimilarity measure consists of different dissimilarity scores for different attributes, which are denoted by δ_{ijk}.

In the remaining of the section, we introduce the dissimilarity measures considered in this paper, that are, the Euclidean Distance, the Hamming Distance, the Heterogeneous Euclidean-Overlap Metric, and the Adapted Gower Coefficient. Since we will not consider missing values in this paper, we do not describe the way they handle missing values here, although both the Heterogeneous Euclidean-Overlap Metric and the Adapted Gower Coefficient originally have built-in ways to handle missing values (see [14, 30]).

3.1 Euclidean Distance

Probably, the best known dissimilarity measure is the Euclidean Distance. When we assume that all attributes are numerical, we can define the Euclidean Distance as

$$\delta_{ij} = \sqrt{\sum_{k=1}^{K} (x_{ik} - x_{jk})^2} \ . \tag{1}$$

However, in this form, the measure is very sensitive to the scale of the different attributes. Therefore, attributes should be normalized first. The most natural way to do this is by computing Z-scores

$$z_{ik} = \frac{x_{ik} - \mu_k}{\sigma_k} \ , \tag{2}$$

where μ_k and σ_k are the mean and standard deviation of attribute k respectively.

The most common way to include (multi-valued) categorical attributes in the Euclidean Distance is by representing every category by a dummy variable. Also, these dummy variables should be normalized to Z-scores. However, to ensure that categorical attributes with a lot of categories will not dominate the dissimilarity, these attributes are normalized. When we define $(x_{ik1}, \ldots, x_{ikL_k})$ as the L_k dummies that describe attribute k for product i (note that $L_k = 1$ for numerical attributes) and incorporate weights w_k, the normalized Euclidean

Distance is defined as

$$\delta_{ij} = \sqrt{\sum_{k=1}^{K} \left(w_k L_k^{-1} \sum_{\ell=1}^{L_k} \left(\frac{x_{ik\ell} - x_{jk\ell}}{\sigma_k \ell} \right)^2 \right)} \ . \tag{3}$$

Note that we can rewrite the Euclidean Distance in terms of dissimilarity scores in the following way

$$\begin{aligned} \delta_{ij} &= \sum_{k=1}^{K} w_k \delta_{ijk} & (4) \\ \delta_{ijk} &= L_k^{-1} \sum_{\ell=1}^{L_k} \left(\frac{x_{ik\ell} - x_{jk\ell}}{\sigma_k \ell} \right)^2 & (5) \end{aligned}$$

3.2 Hamming Distance

On the other hand, when one only has categorical attributes, the Hamming Distance is the best known option. The Hamming Distance is defined as

$$\delta_{ij} = \sum_{k=1}^{K} w_k 1(x_{ik} \neq x_{jk}) \ , \tag{6}$$

where 1() is the indicator function returning a value of 1 when the condition is true and 0 otherwise. Hence, the Hamming Distance counts the number of attributes on which both products disagree. Incorporating numerical attributes directly into this Hamming Distance is therefore not always a good idea, since two numerical values will almost always be unequal to each other. A solution for this might be to discretize the numerical attribute by creating bins containing an equal portion of products. Note that information is lost in this way, since the distance between two bins can only be 0 or 1. However, in paper we consider the Hamming Distance in its most simple form.

Both the Euclidean and Hamming Distance are used as simple reference measures with which we will compare the more complex measures discussed later.

3.3 Heterogeneous Euclidean-Overlap Metric

The Heterogeneous Euclidean-Overlap Metric (HEOM) [30] has been used in a recommender system setting in [3, 25, 26]. HEOM computes the dissimilarity scores for categorical and numerical attributes in two different ways.

The dissimilarity score for numerical attributes is computed using

$$\delta_{ijk} = \frac{|x_{ik} - x_{jk}|}{\max(\mathbf{x}_k) - \min(\mathbf{x}_k)} \ , \tag{7}$$

where $\min(\mathbf{x}_k)$ and $\max(\mathbf{x}_k)$ are the minimum and maximum value of attribute k in the catalog, such that δ_{ijk} is always in between 0 and 1.

For categorical attributes the overlap is computed, as is done by the Hamming Distance

$$\delta_{ijk} = 1(x_{ik} \neq x_{jk}) \ . \tag{8}$$

Finally, the dissimilarity scores are combined

$$\delta_{ij} = \sqrt{\sum_{k=1}^{K} w_k \delta_{ijk}^2} \ . \tag{9}$$

Since HEOM does not have a built-in way to handle multi-valued categorical attributes, these are treated as regular categorical attributes, where each unique combinations of values is treated as a category.

3.4 Adapted Gower Coefficient

The Adapted Gower Coefficient (AGC), introduced in [14] as an adaptation of the General Coefficient of Similarity proposed by Gower [11], is, similar to HEOM, also based on different specifications of dissimilarity scores for different types of attributes. Both measures mainly differ in the way how dissimilarity scores are normalized.

For numerical attributes the dissimilarity score is computed in the following way in the AGC framework

$$\delta_{ijk} = \frac{|x_{ik} - x_{jk}|}{((n^2-n)/2)^{-1} \sum_{i<j} |x_{ik} - x_{jk}|} \ , \tag{10}$$

which is based on the absolute distance normalized in such a way that the dissimilarity scores have an average of 1 in the product catalog for each attribute k. For categorical attributes, the same approach has been followed, only now using the overlap metric

$$\delta_{ijk} = \frac{1(x_{ik} \neq x_{jk})}{((n^2-n)/2)^{-1} \sum_{i<j} 1(x_{ik} \neq x_{jk})} \ . \tag{11}$$

To be able to handle multi-valued categorical attributes in the AGC framework, we use the approach introduced in [13], which assumes that two products are identical on a multi-valued categorical attribute, when they share exactly the same values. So, in AGC the dissimilarity score for a multi-valued categorical attribute is determined by counting the number of values that only one of the products has. More formally, the dissimilarity score δ_{ijk} for multi-valued categorical attributes is defined as

$$\delta_{ijk} = \frac{|x_{ik} \Delta x_{jk}|}{((n^2-n)/2)^{-1} \sum_{i<j} |x_{ik} \Delta x_{jk}|} \ , \tag{12}$$

where both x_{ik} and x_{jk} are sets of values and Δ is the symmetric difference set operator.

Finally, the dissimilarity scores are combined as follows

$$\delta_{ij} = \sqrt{\sum_{k=1}^{K} w_k \delta_{ijk}} \ . \tag{13}$$

3.5 Estimation of Weights

An important issue in attribute-based dissimilarity measures is the weighting of attributes. In most domains, users do not consider all attributes to be equally important and, therefore, giving all attributes an equal weight will not provide recommendations that are considered to be the best by users. A possible solution is asking the users to specify these weights. However, users will do this without taking into account specific issues of the dissimilarity measure used.

In our user experiment described in the next section, we create a relative relevance matrix $\mathbf{R}$ in which r_{ij} describes how relevant product j is as a recommendation when looking at reference product i, where a value of 0 means that j is considered by no one to be relevant and a value of 1 means that everyone considered j to be relevant. Good recommendations are recommendations considered to be relevant by many users and, therefore, this matrix can be seen as an ideally specified similarity matrix. In other words, a similarity matrix providing the best recommendations is a similarity matrix that is (almost) identical to $\mathbf{R}$.

We can use this idea to estimate attribute weights in the following way. Since, we use dissimilarity measures and not similarity measures we first transform the relevative relevances in $\mathbf{R}$ to a measure of dissimilarity, that is, we introduce the optimal dissimilarities $\delta^*_{ij} = 1 - r_{ij}$. We are interested in finding weights w_k that combined with dissimilarity scores δ_{ijk} provided by one of the dissimilarity measures discussed the previous section lead to dissimilarities best matching the optimal dissimilarities δ^*_{ij}. When having a dissimilarity measure for which it holds that

$$\delta_{ij} = \sum_{k=1}^{K} w_k \delta_{ijk} \tag{14}$$

we can estimate these weights using ordinary least squares regression. We first collect all nondiagonal elements of the optimal dissimilarity matrix in a $(n^2 - n)$-length vector $\mathbf{o}$. Correspondingly, we create for all attributes such vectors consisting of dissimilarity scores and combine these vectors in $(n^2-n) \times K$ matrix $\mathbf{D}$. Hence, a row of $\mathbf{D}$ contains dissimilarity scores δ_{ijk} for a certain attribute k between all pairs ij. Regression coefficients $\mathbf{b}$ can then be determined using

$$\mathbf{b} = (\mathbf{D}'\mathbf{D})^{-1}\mathbf{D}\mathbf{o} \ . \tag{15}$$

For ease of interpretability we can transform these regression coefficients such that they sum up to one (without changing performance of the dissimilarity measure in our evaluation)

$$w_k = \frac{b_k}{\sum_{k'=1}^{K} b'_k} \ . \tag{16}$$

However, not all dissimilarity measures used in this paper are in the form of (14). In such cases transformations are necessary. For HEOM, for example, we define $\mathbf{D}$ to contain the squared dissimilarity scores to match (9). Since also a square root is taken in (9), we let also $\mathbf{o}$ contain the squared optimal dissimilarities. The latter is also done for the Euclidean distance and the Adapted Gower Coefficient.

Note that weights can also become negative in this procedure, which may be counterintuitive. Therefore, it might be the case that better weights are obtained when using nonnegative least squares.

Table 1: Characteristics of the seven attributes for the 25 selected microwave ovens

Numerical Attributes				
Name	Mean	Standard Deviation	Minimum	Maximum
Price (€)	606.82	455.87	134	1988
Volume (liters)	36.68	7.72	24	50
Max. Power (Watt)	906.00	50.67	800	1000
Categorical Attributes				
Name	Values (number of occurrences)			
Brand	Siemens (8), Sharp (5), Whirlpool (4), Samsung (3), and 5 other brands (5)			
Type	Combi (23), Combi with Steam (2)			
Model	Built-In (15), Countertop (10)			
Multi-Valued Categorical Attributes				
Name	Values (number of occurrences)			
Color	Steel Grey (16), Aluminium (3), Silver (3), White (3), and 2 other colors (2)			

4 Evaluation Methodology

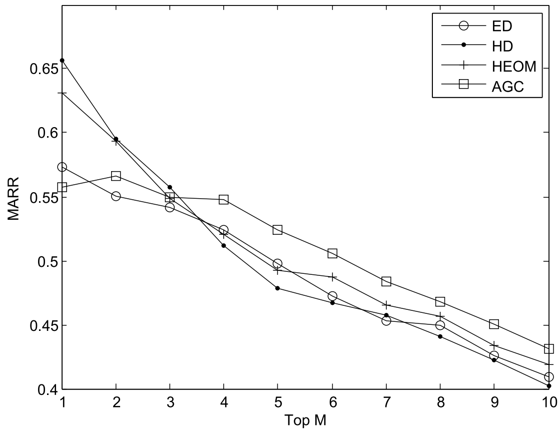
In contrast to other evaluation studies of recommender systems, our approach is not based on user behavior on a website, but on an online survey we conducted. People were asked to participate in this survey via the Vergelijk.nl newsletter. Vergelijk.nl is one of the leading price comparison sites in the Netherlands. In total, 70 people completed the survey.

In the survey, we used a product catalog of microwave ovens provided by Compare Group, the owner of Vergelijk.nl. This catalog consists of 236 microwave ovens described by 30 attributes. However, we have decided to use only the 25 most popular microwave ovens available in this product catalog, such that we have observations for each reference/suggestion combination. Also, we showed only seven attributes to the users (the attributes having no missing values in the complete product catalog) and, therefore, only use these seven attributes in our evaluation. Characteristics of these attributes in the selected data set are shown in Table 1.

In this survey, we presented the respondents three times a reference product, which they were asked to consider as the product they were looking at right now on some website. On the same page we presented them six randomly selected products, which they should consider as possible recommendations given that they were looking at the reference product. Then, respondents were asked to select the products they considered to be relevant recommendations. In total, respondents were given three different reference products and for each of these reference products we gave them twelve possible recommendations. Both the reference products and possible recommendations were selected at random. At the end of the survey, we also asked the users how important they found these seven attributes on a 5 point scale.

4.1 Mean Average Relative Relevance

We created two matrices based on the results from the survey. In the first matrix **A** we counted for each product j how often it was considered to be relevant given that product i was the reference product. For the second matrix **B** we counted for each product j how often it was shown to respondents given that product i was the reference product. Given these two matrices, we can compute the relative relevance matrix **R** by computing $r_{ij} = a_{ij}/b_{ij}$ for every pair ij. Note that **R** is a 25×25 matrix of which the diagonal does not contain



Top	1	3	5	10
ED	0.573	0.542	0.498	0.409
HD	0.656	0.557	0.479	0.402
HEOM	0.632	0.548	0.493	0.419
AGC	0.557	0.550	0.524	0.431

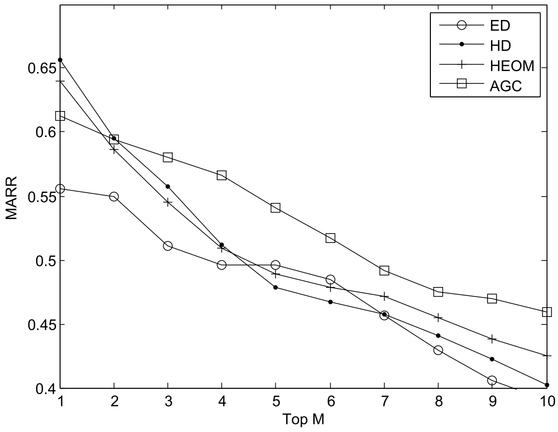
Figure 1: MARR of the four dissimilarity measures using equal attribute weights.

relevant data, since a reference product cannot be a relevant recommendation for itself.

This matrix $\mathbf{R}$ can be used to evaluate a dissimilarity measure in the following way. First we compute dissimilarity matrix $\boldsymbol{\Delta}$ using the same products as in $\mathbf{R}$ and the dissimilarity measure of our choice. Given this matrix $\boldsymbol{\Delta}$ we can determine for each reference product i a ranking of recommended products by sorting row i of $\boldsymbol{\Delta}$. Let vector $\mathbf{r}_i^*$ be the 24×1 vector of relative relevances for reference product i ordered according to the ranking defined by $\boldsymbol{\Delta}$. Then, we can compute for a given number of recommendations (a Top M) a measure we call the mean average relative relevance (MARR) in the following way

$$\begin{aligned} MARR_M &= \frac{1}{nM} \sum_{i=1}^{n} \sum_{m=1}^{M} r_{im}^* \qquad (17) \\ &= \frac{1}{nM} \sum_{i=1}^{I} \sum_{m=1}^{M} \frac{a_{i,\psi_i(m)}}{b_{i,\psi_i(m)}} \ , \end{aligned}$$

where $\psi_i(m)$ is the column index of the m-th smallest dissimilarity in row i of dissimilarity matrix $\boldsymbol{\Delta}$ excluding element ii. So, for each product i, the average relative relevance is computed over the M products that rank highest in similarity with product i according to the dissimilarity measure we want to evaluate. Note that the matrix $\mathbf{R}$ in our experiment (using 25 products and having 70 respondents) is based on, on average, 4.2 reference/suggestions combinations per cell.



Top	1	3	5	10
ED	0.555	0.511	0.496	0.391
HD	0.656	0.557	0.479	0.402
$HEOM$	0.640	0.545	0.489	0.425
AGC	0.612	0.580	0.541	0.460

Figure 2: MARR of the four dissimilarity measures normalized on the complete product catalog and using equal attribute weights.

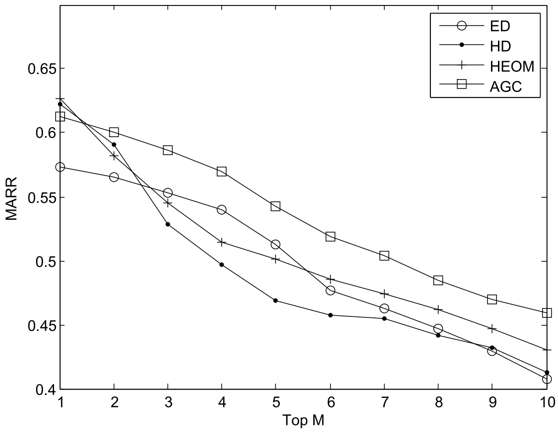
5 Evaluation Results

We have used the MARR measure to evaluate the performance of the dissimilarity measures and show results of an evaluation using equal weights in Figure 1. Normalization of the dissimilarity measures was done only using the 25 selected products.

In this figure and during the remainder of this paper, we use the following abbreviations: *ED* for the Euclidean Distance, *HD* for the Hamming Distance, *HEOM* for the Heterogeneous Euclidean Overlap Measure, and AGC for the Adapted Gower Coefficient.

These results show that when only recommending a single product the Hamming distance provides the most relevant recommendations. Conversely, AGC performs poor when recommending a single product, but performs best when recommending a group of 5 or more products.

All dissimilarity measures except for the *HD* use some form of normalization, namely the Z-scores in the *EC*, the division by range for numerical attributes in *HEOM*, and the normalization of dissimilarity scores in the AGC measure. Since the selection of the 25 microwave ovens is based on product popularity, these microwave ovens are not a representative selection for the complete product catalog and, therefore, performance of these dissimilarity measures may be improved when normalizing over the complete product catalog. MARR results using normalization and equal weights are shown in Figure 2. Note that the *HD* used here is identical to the one used before, since no normalization is required to compute the *HD*. The normalization based on the complete data set is beneficial to AGC, although it does still worse than *HD* and *HEOM* when recommending only a single product. For *HEOM* there is not much difference in results and normalization based on the complete product catalog seriously weakens performance for *ED*. Since normalizing on the complete data set seems



Top	1	3	5	10
ED	0.573	0.553	0.513	0.408
HD	0.622	0.529	0.469	0.413
HEOM	0.618	0.543	0.507	0.430
AGC	0.612	0.586	0.543	0.459

Figure 3: MARR of the four dissimilarity measures using average attribute importance given by respondents as attribute weights.

to be beneficial to AGC, we use this approach for this dissimilarity measure in the remainder of the paper. The other dissimilarity measures are normalized on the 25 selected products.

5.1 Evaluating Attribute Weights

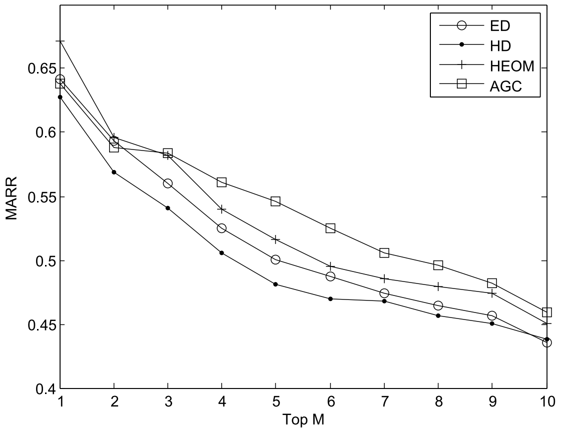
Until now, we analyzed the dissimilarity measures assuming that all attributes are considered to be equally important by users. However, in the survey we also asked the respondents how important they found the attributes used in this study. When use the average importance stated by users to determine weights, we get the weights shown in the column named Importance in Table 2. Using these weights in combination with the four dissimilarity measures, we get the results shown in Figure 3.

Using these weights leads to small improvements of performance for ED and AGC, while the results for the other two dissimilarity measures are worse than when using equal weights. We see three possible reasons why using these weights does not work very well. The first reason might be that respondents act different than they think they do, that is, although they say some attribute is quite important, they do not actually use this attribute in determining why a product is relevant or not. Second, respondents might normalize attributes differently than the dissimilarity measures do, which might lead to wrong weights. Third, since all importances are relatively close to each other, the weighting used might not be strong enough and better results might be obtained by, for instance, using the squared importance as weight.

In Section 3.5, we discussed how we could estimate weights that are optimal given the relative relevance matrix and the dissimilarity scores provided by a dissimilarity measure. Table 2 shows these weights that resulted from this method for the four dissimilarity measures. As can be seen these weights differ

Table 2: Optimal attribute weights for the four dissimilarity measures. The column Importance shows the normalized weights based on the average importance stated by users in the survey.

Attribute	Importance	*ED*	*HD*	*HEOM*	*AGC*
Price	0.166	0.094	0.330	0.087	0.145
Volume	0.158	0.207	0.236	0.421	0.357
Max. Power	0.165	0.038	0.059	0.034	-0.056
Brand	0.124	0.233	0.089	0.112	0.138
Type	0.127	0.002	-0.071	-0.008	0.001
Model	0.138	0.236	0.285	0.296	0.326
Color	0.120	0.190	0.072	0.059	0.089

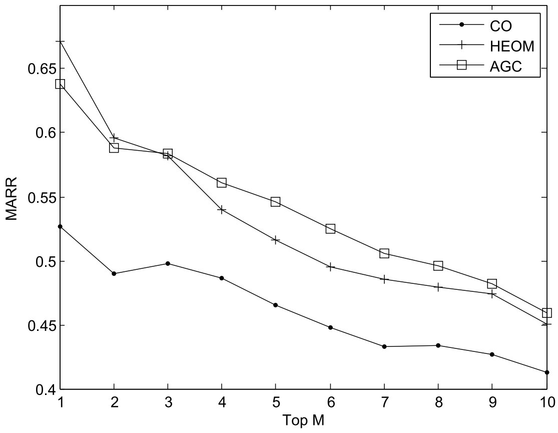


Top	1	3	5	10
ED	0.641	0.560	0.500	0.435
HD	0.628	0.541	0.481	0.438
HEOM	0.672	0.582	0.516	0.450
AGC	0.638	0.584	0.546	0.459

Figure 4: MARR of the four dissimilarity measures using optimal attribute weights.

quite a lot from the weights given by users and among the dissimilarity measures. For example, Max. Power was according to the users one of the most important attributes, while it has received very low weights for all dissimilarity measures. A reason might be, that the 25 selected microwave ovens do not differ much on their maximum power, which also holds for their type. On the other hand, Volume and Model are attributes that received very high weights for all dissimilarity measures.

Figure 4 shows the performance of the dissimilarities using these optimal weights. Except for the Hamming Distance all dissimilarity measures show an improvement. When recommending a single product *HEOM* performs best, while when recommending three products *HEOM*, and *AGC* show similar performance. If one wants to recommend more than three products *AGC* outperforms *HEOM* and this measure should be preferred. It is remarkable that the Euclidean distance seems to benefit most from the optimal weighting procedure.



Top	1	3	5	10
CO	0.526	0.498	0.465	0.413

Figure 5: MARR of the recommendation approach based on co-occurrences (CO). MARR of other dissimilarity measures is equal to the ones shown in Figure 4.

5.2 Comparison with Collaborative Method

Both in science and practice, collaborative recommendation methods [10], are more popular than content-based methods. Main advantage of these methods is that they only rely on co-occurrences (such as co-purchases and co-viewings) and, thus, no specific product data or domain knowledge is required. On the other hand, collaborative methods have two limitations in recommendation of durable goods. First, collaborative methods lack the ability to recommend new or recently introduced products, since no or insufficient data is available in such cases. Second, they can only recommend products given a product available in the product catalog and not based on an ideal product specification or a search query specified by the users. Case-based approaches do not have these limitations.

We compared the dissimilarity measures with a simple approach to collaborative filtering. In a clicksteam log of the Vergelijk.nl website of the period July 15 until September 15, 2007, we counted co-occurrences of the 25 microwave ovens. A co-occurrence is defined as a session in which of both microwave ovens the details page was visited. When a reference product is given, we simply order the products on the number of co-occurrences with this reference product to retrieve a recommendations list. Based on these recommendation lists, we can compute the MARR.

We compared the co-occurrence approach to the dissimilarity measures using optimal weights in Figure 5. As can be seen in this plot, the co-occurrence approach performs much worse than these, but also all other, dissimilarity measure based approaches. It seems that counting co-occurrences of products in sessions on a website is not a good method for providing relevant recommendation. A reason for this might be that co-occurrences might for a substantial part be determined by the structure of the website. Products located just above or below the reference product have a larger probability to co-occur with this product, although they might be quite different.

6 Conclusions

In this paper, we evaluated four dissimilarity measures based on an online survey in which we asked people to select relevant recommendations given a reference product. Based on the survey responses, we created a relative relevance matrix and used this matrix to compute the mean average relative relevance (MARR) of recommendation lists of four dissimilarity measures.

We used a selection of 25 microwave ovens and seven attributes from a microwave oven product catalog in our evaluation. When we considered all attributes to be equally important, the Hamming Distance performed best when a single recommendation was evaluated. For top 3's and higher, the Adapted Gower Coefficient is the best alternative. We also evaluated the dissimilarity measures using weights based on the average importance of attributes as stated by the respondents. This did not lead to much better results.

Also, another weighting approach was evaluated, in which the weights were determined using linear regression based on dissimilarity scores and the relative relevance matrix. Weights derived using this method differed quite a lot from the weights specified by the respondents. This may be the case due to the selected products, which might have some bias, but also due to the specific properties of the dissimilarity measures, which are unknown to the respondents. Using these weights, performance of three of the four dissimilarity measures was improved. When recommending a single product, the Heterogeneous Euclidean-Overlap Metric performed best and we advise to use this measure for this purpose. If one would like to recommend more than a single product, the Adapted Gower Coefficient should be preferred. Finally, we showed that all these dissimilarity approaches outperform a simple collaborative filtering approach.

Future research should focus on validating these results on other product catalogs and in more complex situations. We limited the number of attributes and number of products. Although this makes the tasks given to respondents manageable and it is more easy to fill the relative relevance matrix, this is not a real life situation. Also, missing values were not considered in this study. Another challenge is to collect this kind of data in an operational recommender system. Nevertheless, we showed in this paper a new promising evaluation approach for recommender systems that is independent of the website structure.

Acknowledgements

We thank Compare Group for making their product catalog and clickstream data available to us and for distributing our survey via the Vergelijk.nl newsletter.

References

[1] A. Aamodt and E. Plaza. Case-based reasoning: Foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59, 1994.

[2] G. Adomavicius and A. Tuzhilin. Towards the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *IEEE Trans. Knowl. Data Eng.*, 17(6):734–749, 2005.

[3] B. Arslan, F. Ricci, N. Mirzadeh, and A. Venturini. A dynamic approach to feature weighting. *Manage. Inform. Syst.*, 6:999–1008, 2002.

[4] D. Billsus, M. J. Pazzani, and J. Chen. A learning agent for wireless news access. In *Proceedings of the 5th International Conference on Intelligent User Interfaces*, pages 33–36, 2000.

[5] L. K. Branting. Learning feature weights from customer return-set selections. *Knowl. Inf. Syst.*, 6:188–202, 2004.

[6] J. S. Breese, D. Heckerman, and C. Kadie. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the 14th Annual Conference on Uncertainty in Artificial Intelligence*, pages 43–52, 1998.

[7] R. Burke. Knowledge based recommender systems. *Encyclopedia of Library and Information Science*, volume 69, Supplement 32. Marcel Dekker, New York, 2000.

[8] R. Burke. Hybrid recommender systems: Survey and experiments. *User Model. User-Adapt. Interact.*, 12:331–370, 2002.

[9] L. Coyle and P. Cunningham. Improving recommendation rankings by learning personal feature weights. *Lect. Notes Comput. Sci.*, 3155:560–572, 2004.

[10] D. Goldberg, D. Nichols, B. M. Oki, and D. Terry. Using collaborative filtering to weave an information tapestry. *Comm. ACM*, 35(12):61–70, 1992.

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

[12] J. L. Herlocker, J. A. Konstan, and J. T. Riedl. An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Inf. Retr.*, 5:287–310, 2002.

[13] M. Kagie, M. van Wezel, and P. J. F. Groenen. Choosing attribute weights for item dissimilarity using clickstream data with an application to a product catalog map. In *Proceedings of the 2nd ACM Conference on Recommender Systems*, pages 195–202, 2008.

[14] M. Kagie, M. van Wezel, and P. J. F. Groenen. A graphical shopping interface based on product attributes. *Decis. Support Syst.*, 46(1):265–276, 2008.

[15] M. Kagie, M. Van Wezel, and P. J. F. Groenen. Determination of attribute weights for recommender systems based on product popularity. Technical report, Erasmus Research Institute in Management, Erasmus University Rotterdam, 2009. forthcoming.

[16] J. W. Kim, B. H. Lee, M. J. Shaw, H.-L. Chang, and M. Nelson. Application of decision-tree induction techniques to personalized advertisements on internet storefronts. *Int. J. Electron. Commer.*, 5(3):45–62, 2001.

[17] T. W. Liao, Z. Zhang, and C. R. Mount. Similarity measures for retrieval in case-based reasoning systems. *Appl. Artif. Intell.*, 12:267–288, 1998.

[18] G. Linden, B. Smith, and J. York. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Comput.*, 7(1):76–80, 2003.

[19] F. Lorenzi and F. Ricci. Case-based recommender systems: A unifying view. *Lect. Notes Comput. Sci.*, 3169:89–113, 2005.

[20] D. O'Sullivan, B. Smyth, and D. Wilson. Understanding case-based recommendation: A similarity knowledge perspective. *Int. J. Artif. Intell. Tools*, 14(1–2):215–232, 2005.

[21] M. J. Pazzani and D. Billsus. Learning and revising user profiles: The identification of interesting web sites. *Mach. Learn.*, 27:313–331, 1997.

[22] M. J. Pazzani and D. Billsus. Content-based recommendation system. *Lect. Notes Comput. Sci.*, pages 4321:325–341, 2007.

[23] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. T. Riedl. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work*, pages 175–186, 1994.

[24] P. Resnick and H. R. Varian. Recommender systems. *Comm. ACM*, 40(3):56–58, 1997.

[25] F. Ricci and F. Del Missier. Supporting travel decision making throug personalized recommendation. In *Designing Personalized User Experiences in eCommerce*, volume 5 of *Human-Computer Interaction Series*, chapter 4, pages 231–251. Springer, Netherlands, 2004.

[26] J. Sandvig and R. Burke. Aacorn: A CBR recommender for academic advising. Technical Report TR05-15, DePaul University, 2005.

[27] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. T. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International Conference on World Wide Web*, pages 285–295, 2001.

[28] 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, 2000.

[29] D. Wettschereck and D. W. Aha. Weighting features. *Lect. Notes Comput. Sci.*, 1010:347–358, 1995.

[30] D. R. Wilson and T. R. Martinez. Improved heterogeneous distance functions. *J. Artif. Intell. Res.*, 6:1–34, 1997.

[31] M. Zanker, M. Jessenitschnig, D. Jannach, and S. Gordea. Comparing recommendation strategies in a commercial context. *IEEE Intelligent Systems*, 22(3):69–73, 2007.

Publications in the Report Series Research[*] in Management

ERIM Research Program: “Marketing”

2009

Map Based Visualization of Product Catalogs
Martijn Kagie, Michiel van Wezel, and Patrick J.F. Groenen
ERS-2009-010-MKT
http://hdl.handle.net/1765/15142

Embedding the Organizational Culture Profile into Schwartz’s Universal Value Theory using Multidimensional Scaling with Regional Restrictions
Ingwer Borg, Patrick J.F. Groenen, Karen A. Jehn, Wolfgang Bilsky, and Shalom H. Schwartz
ERS-2009-017-MKT
http://hdl.handle.net/1765/15404

Determination of Attribute Weights for Recommender Systems Based on Product Popularity
Martijn Kagie, Michiel van Wezel, and Patrick J.F. Groenen
ERS-2009-022-MKT
http://hdl.handle.net/1765/15910

An Empirical Comparison of Dissimilarity Measures for Recommender Systems
Martijn Kagie, Michiel van Wezel, and Patrick J.F. Groenen
ERS-2009-023-MKT
http://hdl.handle.net/1765/15911

[*] 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