Similarity Assessment as a Dual Process Model of Counting and Measuring

Bert Klauninger and Horst Eidenberger
Institute for Interactive Media Systems, Vienna University of Technology, Vienna, Austria

Keywords: Dual Process Model, Similarity Assessment, Combined Kernels, Counting and Measuring, Quantisation, Generalisation, Taxonomic and Thematic Reasoning, Image Similarity.

Abstract: Based on recent findings from the field of human similarity perception, we propose a dual process model (DPM) of taxonomic and thematic similarity assessment which can be utilised in machine learning applications. Taxonomic reasoning is related to predicate based measures (counting) whereas thematic reasoning is mostly associated with metric distances (measuring). We suggest a procedure that combines both processes into a single similarity kernel. For each feature dimension of the observational data, an optimal measure is selected by a Greedy algorithm: A set of possible measures is tested, and the one that leads to improved classification performance of the whole model is denoted. These measures are combined into a single SVM kernel by means of generalisation (converting distances into similarities) and quantisation (applying predicate based measures to interval scale data). We then demonstrate how to apply our model to a classification problem of MPEG-7 features from a test set of images. Evaluation shows that the performance of the DPM kernel is superior to those of the standard SVM kernels. This supports our theory that the DPM comes closer to human similarity judgment than any singular measure, and it motivates our suggestion to employ the DPM not only in image retrieval but also in related tasks.

1 INTRODUCTION

We suggest an SVM kernel function for image retrieval that is based on the latest findings of psychological research on human similarity measurement. Humans appear to derive their similarity judgments from a mixture of *thematic* and *taxonomic* stimuli, called a *dual process model* of similarity (DPM) (Wisniewski and Bassok, 1999). Thematic stimuli (*e.g.*general appearance in form of a global color histogram) are typically measured by distance functions, taxonomic ones (*e.g.*co-existing properties such as "person visible") by similarity functions. The first require the transformation from distance to similarity by a so-called *generalisation* function. The latter require the *quantisation* of numbers into predicates.

This paper provides the DPM kernel as well as the necessary components, partly taken from earlier work of the authors. The kernel is applied on a set of MPEG-7 features computed for a test set of images. Evaluation shows that for the present task the DPM kernel is superior to the standard SVM kernels which proofs the concept as well as supports the mentioned psychological findings. We suggest to employ the DPM kernel also on related tasks.

The next section introduces the model components, including the generalisation model and the quantisation model employed in the DPM kernel. Section 3 explains the similarity model itself, evaluated and discussed in Section 4.

2 BACKGROUND

2.1 Measuring vs. Counting

Traditional mathematical models of human similarity perception are often geometric ones (*e.g.*(Torgerson, 1952)): Observational data is mapped into some psychological or mathematical space (usually a vector space) in such a way that similar observations are projected into the same regions, thus constituting a *metric* on that space.

In his 1977 paper (Tversky, 1977), psychologist Amos Tversky proposed a different approach towards similarity. After demonstrating that the metric axioms are in fact violated in human similarity judgment, he

Klauninger, B. and Eidenberger, H.
Similarity Assessment as a Dual Process Model of Counting and Measuring.
DOI: 10.5220/0005655801410147
In *Proceedings of the 5th International Conference on Pattern Recognition Applications and Methods (ICPRAM 2016)*, pages 141-147
ISBN: 978-989-758-173-1

developed a set-theoretical framework that operates in terms of matching and mismatching features rather than metric distances. Observations are represented as collection of features, and similarity is computed by counting common and distinct features. Figure 1

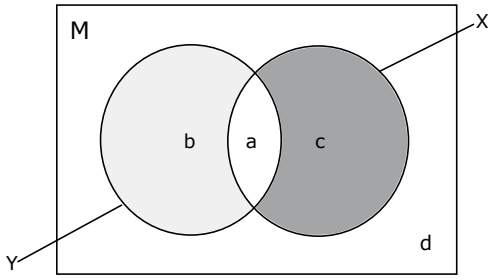

Figure 1: Feature Contrast Model.

defines the building blocks of all predicate based similarity measures: For two stimuli X, Y : $a = |X \cap Y|$, $b = |X \setminus Y|$, $c = |Y \setminus X|$, and $d = |M \setminus (X \cup Y)|$. The concept of similarity as counting gave birth to a variety of *predicate based measures*, as for example Tversky's *Feature Contrast Model* (FCM) itself, Jaccard's coefficient, Kulczynski distance and many others. Other existing measures were integrated in the model (*e.g.*the Hamming distance). Both metric distances and predicate-based measures are essential building blocks of the DPM kernel defined in Section 3.

2.2 Quantisation Model

In order to be able to apply predicate-based measures to quantities, the set theoretical definition of predicates has to be extended to the interval scale, i.e. each feature dimension indicates that a property is more or less present (0 means "not at all", 1 means "fully present"). One candidate would be the *Fuzzy Feature Contrast Model* (see (Santini and Jain, 1999)) which arguably suffers from some shortcomings. So, for our purposes, we utilized the Quantisation Model (*QM*) as described in (Eidenberger, 2003). This QM replaces set theoretical predicate measures with statistical functions in the following way:

$$a = |X \cap Y| := \sum_i s_i, s_i = \begin{cases} \frac{x_i + y_i}{2} & \text{if } M - \frac{x_i + y_i}{2} <= \varepsilon_1 \\ 0 & \text{else} \end{cases} \quad (1)$$

$$b = |X \setminus Y| := \sum_i s_i, s_i = \begin{cases} x_i - y_i & \text{if } M - (x_i - y_i) <= \varepsilon_2 \\ 0 & \text{else} \end{cases} \quad (2)$$

$$c = |Y \setminus X| := \sum_i s_i, s_i = \begin{cases} y_i - x_i & \text{if } M - (y_i - x_i) <= \varepsilon_2 \\ 0 & \text{else} \end{cases} \quad (3)$$

with: $\varepsilon_1, \varepsilon_2$ being two thresholds and $M = 1$ for $x_i \in [0,1]$. Depending on the thresholds, two quantities are considered co-existing predicates if they are both sufficiently large. Below, we employ this model for the transformation of image features into countable taxonomic properties.

2.3 From Distance to Similarity: Generalisation

Intuitively, one would define similarity as some kind of "inverse distance": The more similar two stimuli are, the smaller their distance in an appropriate psychological space should be. A function that estimates similarities from distances is called a *generalisation function*.

In (Shepard, 1987), Shepard carefully examined different candidates and came to the conclusion that the probability $P(R_x|S_y)$ that a stimulus S_y is associated with a response R_x is proportional to $e^{-\delta_{x,y}}$, δ being the distance between representations x and y in an appropriate space (Shepard, 1987). This relation is known as the *Universal Law of Generalisation*.

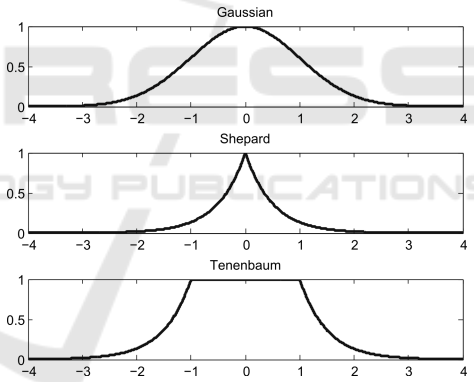

Figure 2: Three Generalisation Functions.

While Shepard's law holds on some classes of data sets, there are others that violate it (Chater and Vitnyi, 2003, p.349). In some cases, a *Gaussian* density function, where $P(R_x|S_y)$ is proportional to $e^{-\delta_{x,y}{}^2}$, has been applied more successful. Another generalisation function has been proposed by Tenenbaum and Griffiths (Tenenbaum and Griffiths, 2001) as an extension of the Universal Law to generalising from multiple consequential stimuli. In the DPM kernel, we employ a generalisation function to transform distance values into thematic similarities that can be combined with taxonomic properties.

2.4 General Dual Process Model

For the definition of our DPM kernel function, we employ the simple model suggested in (Eidenberger, 2012, p.540):

$$m_{dpm} = \alpha.m_{tax} + (1-\alpha)g(m_{them}) \quad (4)$$

Here, m_{tax} stands for taxonomic measures which are usually similarities (*e.g.*co-occurrences, cosine similarity); m_{them} stands for the thematic aspects expressed in image features which are usually distances (*e.g.*Hamming distance, city block metric); g is the generalisation function (Gaussian, Tenenbaum or Shepard).

Linear combinations with α are capable of representing any other similarity measurement. α itself is defined by the user's preference toward taxonomic vs. thematic reasoning which could be estimated in psychological tests.

3 MATERIALS AND METHODS

3.1 An Integrative Image of Similarity

Our hypothesis is that, in order to construct an ideal global similarity measure for a given dataset, different feature dimensions may require different similarity functions. The list of possible kernel functions can be aligned at two dimensions (see Figure 3), reflecting current findings from psychology on human similarity perception: the continuum from *similarities* (separable stimuli) to *distances* (integral stimuli) on one hand, and the continuum from *predicate-based* (taxonomic) to *quantitative* (thematic) reasoning on the other hand.

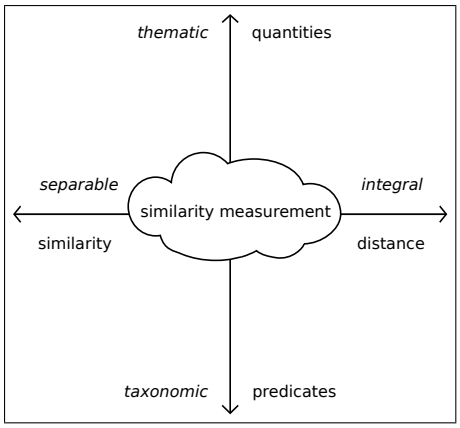

Figure 3: Axes of Similarity Assessment.

3.2 Greedy Algorithm

If we accept the model described above, the question arises how to determine which feature dimensions are best compared by which measure. For a pair of images, we could use some a-priori knowledge about the semantic meaning of the dimension, *e.g.*the feature "global color histogram" would most likely be compared using a quantitative measure (*e.g.*earth mover's distance), whereas for the taxonomic feature "face visible" it would make sense to use some predicate-based measure.

In absence of such a-priori knowledge, we could try out several measures for each dimension, evaluate the model and denote the measure that yields the best global result during the evaluation process. Such a heuristic is known as *Greedy algorithm*: We are able to approximate the globally optimal solution by making locally optimal choices (Cormen et al., 2009, p.424f). In contrast to Dynamic Programming where each sub-problem is optimised separately and then combined, our algorithm evaluates the whole problem in each step.

Ideally, the dimensions should be uncorrelated. In practice, this is usually not the case. Hence the order in which the dimensions were processed had been randomised.

3.3 Our Model

There is a vast variety of similarities, distances, correlations and divergences that have been proposed in literature. A classification into thematic and taxonomic measures can be established in the following way: If the formula contains a contrast term in the formula ($x-y$, $\frac{x}{y}$, $\frac{a}{\cdot}$, $\frac{\cdot}{-a}$, $\frac{\cdot}{b}$ or $\frac{\cdot}{c}$), then it is considered thematic, otherwise taxonomic.

In order to capture the two-dimensional framework of similarities vs. distances and predicates vs. quantities, we focused on four candidates that represent extremes in the continuum of possible measures (see figure 4):

1. Dot Product: separable stimuli, thematic reasoning
2. Number of co-occurrences: separable stimuli, taxonomic reasoning
3. L_1 distance: integral stimuli, thematic reasoning
4. Hamming distance: integral stimuli, taxonomic reasoning

The inner product or dot product of two vectors is equivalent to applying the *cosine similarity* measure on $[0,1]$ normalised data. It compares two vectors in

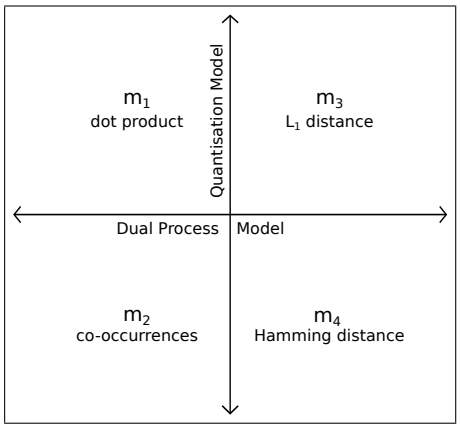

Figure 4: Chosen Measures.

respect of the angle between them: identical vectors result in a score of 1, orthogonal vectors ($\theta = 90°$) give a score of 0. This measure has successfully been applied to fields like document retrieval and data mining (*e.g.*(Salton et al., 1975), (Faloutsos and Lin, 1995)).

In counting the co-occurrences, only those predicates which are present in both sets X and Y are taken into account. Applied to the Quantisation Model (2), this means that those vector components contribute more to the result that have high values in both vectors.

Minkowski distances L_n are appropriate for observations in the form of numeric measurements (*e.g.*measures of geometrical or physical properties). They have been used in models of similarity assessment for a very long time (*e.g.*(Torgerson, 1952)). The *city block distance* L_1 was chosen for its computational effectiveness.

The Hamming distance is defined as the number of different bits in two binary vectors, or set theoretically, as the number of elements that are either present in X *or* in Y, but not in both of them. Quantisation of this measure leads to a distance where large differences in vector components contribute more to the result than small ones.

The dual process Model can now be written as:

$$m_{dpm}(X,Y) = \frac{\alpha}{2}(m_1(X_1,Y_1)+m_2(X_2,Y_2))+ \frac{1-\alpha}{2}g(m_3(X_3,Y_3)+m_4(X_4,Y_4)) \quad (5)$$

with:

$$m_1(\vec{x},\vec{y}) = \vec{x}\cdot\vec{y}$$
$$m_2(\vec{x},\vec{y}) = a(\vec{x},\vec{y})$$
$$m_3(\vec{x},\vec{y}) = L_1(\vec{x},\vec{y}) = \sum_{i=1}^{k}|x_i - y_i|$$
$$m_4(\vec{x},\vec{y}) = b(\vec{x},\vec{y}) + c(\vec{x},\vec{y})$$

X_n and Y_n being the sub-vectors consisting of those components that had been determined to go best with measure m_n by the Greedy algorithm. For generalisation g, the Gaussian function $g(x) = e^{-0.5x^2}$ was selected empirically. a, b and c are computed using the QM (Equations 1, 2 and 3).

Three parameters go into the model: The parameter $\alpha \in [0,1]$ is used to move bias from taxonomic to thematic stimuli, and the parameters ε_i determine the "hardness" of quantisation of the predicate-based measures. In our experiments, α was set to $\frac{k_1}{k}$, k_1 being the number of feature dimensions that are regarded rather separable than integral (by the Greedy algorithm), and k being the total number of dimensions. The quantisation thresholds were tuned heuristically.

4 TESTS AND RESULTS

4.1 Test Set

Our test set consists of 426 instances of MPEG-7 descriptions from images depicting coats of arms. The images are described by 314 dimensions from *CL, CSD, DC, EH, HT, RS, SC* and *TB* descriptors. All components were normalised to [0, 1] beforehand. Ground truth has been provided by manual annotation. Each instance falls into one of five highly semantic categories (*e.g.*"Bavarian city arms") with vastly different sizes (table 1) – to make the machine learning problem harder.

Table 1: Class Sizes.

Class 1	1 instance
Class 2	20 instances
Class 3	34 instances
Class 4	51 instances
Class 5	320 instances

Hence this can be considered a difficult set for training classifiers in general. Another reason for choosing this dataset was the hope that, by design, some features would be more likely to fit into predicate-based than in quantitative measurement. Figure 5 shows some examples of value distributions for different features. The numbers above the histograms denote the corresponding feature dimension. We suspected that features with a continuous spectrum are best compared using quantitative measures whereas features with a small number of sharp peaks can be interpreted as sets of predicates, one for each peak.

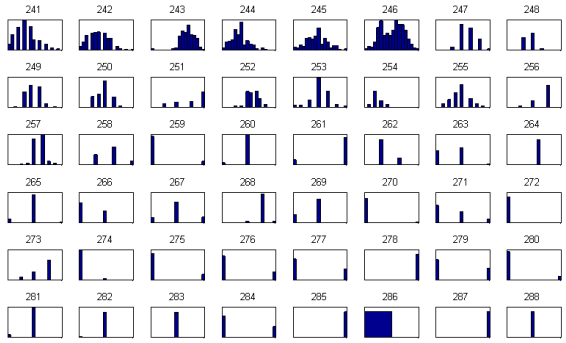

Figure 5: Example of Feature Dimensions with Different Distributions.

4.2 Test Pipeline

In our experiment, classifiers should be trained that predict the class label of an input vector with high confidence. Support Vector Machines (SVM) were chosen because similarity measures can be directly applied as kernels, and because they are fairly efficient in respect to computational cost. In order to extend the SVM concept to a multi-class problem, the approach of one-vs-one (Milgram et al., 2006) was implemented.

Usually, one would reduce the dimensionality, *e.g.*by means of PCA or SVD, but for our experiment, we decided to omit this step in order to keep full diagnosticity for each dimension.

The experimental setting was as follows:

1. Input data (observation vectors and ground truth) is read in and randomly divided into training and test set (using a holdout of 50%).
2. From the training set, for each feature dimension in random order, m_1, m_2, m_3 and m_4 is tried out by the Greedy algorithm: At each step, a set of SVMs is trained one-by-one, using DPM similarity as kernel. The resulting classifier is cross-validated on the test set. From the four possibilities, that measure which yields the best classification result (maximal global F_1 score) is kept for this dimension.
3. In the end, a DPM similarity kernel is obtained that is approximately optimal.
4. The performance of this DPM kernel is compared to single kernels (linear, quadratic, polynomial, Radial Base Function) in terms of *precision, recall, fallout* and F_1 *score*, per class and averaged.

Evaluation was performed in terms of precision, recall, fallout and F_1 score, per class and globally. For global performance estimation of a model, macro-averaging of the per-class results was used as proposed by Yang and Liu, albeit in the context of text categorisation (Yang and Liu, 1999):

$$F_1^{global} = \frac{\sum_{i=1}^{k} F_1(i)}{k} \text{ , k is the number of classes} \quad (6)$$

4.3 Results

Figure 6 shows the precision-recall curves for each class, collected from ten test runs. One figure is given per kernel type. As can be seen, the DPM kernel and the linear kernel come closest to the optimal recall and precision (upper right corner of the figures).

Tables 2 contain the performance indicators, macro-averaged over all classes, hence equally penalising classification error rates among classes of different size.

Table 2: Averaged Global Performance.

	linear	**quadratic**	**polynomial**
avg. precision	0.4598	0.4132	0.3043
avg. recall	0.3914	0.2688	0.2173
avg. fallout	0.1370	0.1704	0.1933
avg. F_1 score	0.4131	0.2829	0.2058

	RBF	**DPM**
avg. precision	0.1509	0.5823
avg. recall	0.2000	0.4446
avg. fallout	0.2000	0.1112
avg. F_1 score	0.1720	0.4836

4.4 Discussion

Our experiments demonstrate that, for our data set of MPEG-7 features, the Dual Process Model as trained by the Greedy algorithm performs always better that the best singular kernel (which was the linear kernel in all cases). In numbers, global average precision of the DPM kernel is 26.66% higher than for the linear kernel, recall is 13.58% higher, fallout is 12.41% lower, and F_1 score is 17.07% higher.

Quadratic and polynomial kernels performed a little worse than the linear one, and RBF had the problem of always associating all instances to the same class, namely the largest one. This is an interesting finding per se. Arguably, RBF by design works well in classification problems where class sizes are in the same order of magnitude, but cannot cope with very unequally sized classes.

In terms of classes, it is obvious that class 1 (only one class member) failed in all models, whereas class 5 (the largest one) performed best with precision, recall and F_1 values close to 1.0. Regarding the smaller classes 2 to 4, the discriminative power of the DPM kernel comes into play. Here the DPM has always performed better in terms of recall, precision and fallout than all other kernels under consideration.

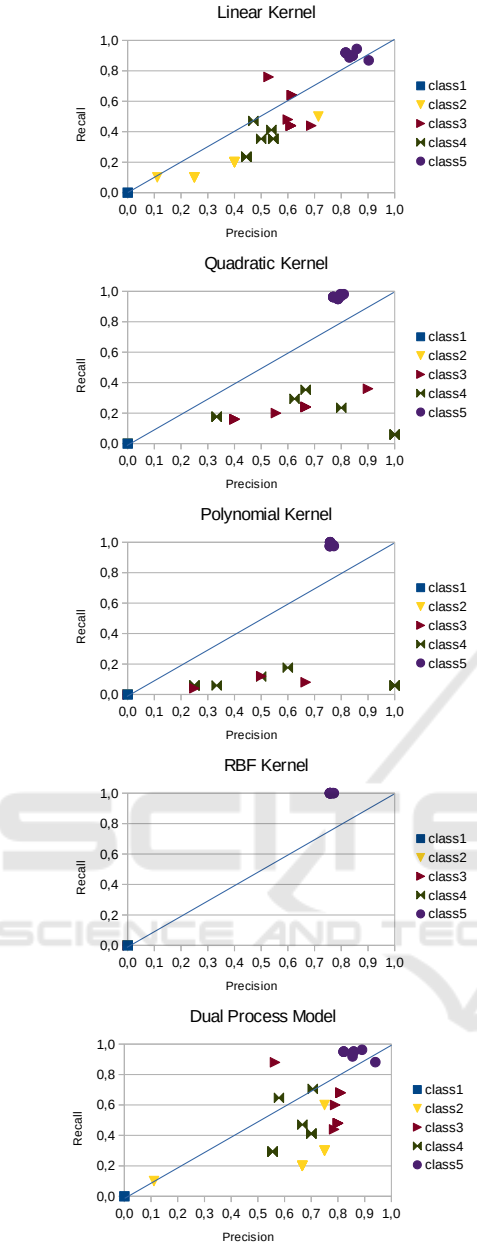

Figure 6: Performance of Different Kernels per Class.

5 CONCLUSIONS AND FUTURE WORK

The experimental results presented in section 4 support the hypothesis that a dual process model of separable and integral stimuli, comprising both geometric and quantised predicate based similarity measures, comes closer to human judgment (ground truth) than the best single measures. The validity of our simple DPM (equation 4) has been demonstrated and thus provided the motivation to direct future research activity towards developing a universal DPM framework for similarity assessment.

For future work, we expect that the value distribution of a feature can give important clues about the kind of similarity measure best applied to it. We will endeavor to identify feature dimensions automatically with histograms consisting only of a small number of peaks, break them up into pseudo-predicates and employ predicate-based measures directly (i.e. without quantisation) on them. Our hope is that a DPM kernel using this approach will come even closer to human similarity assessment while avoiding the costly process of the Greedy algorithm.

The next step will be to apply the DPM to data from different domains in order to proof its universality. We are positive that the development of a universal DPM framework is at reach and will ultimately lead to improved performance in similarity assessment which provides the basis for classification, clustering and information retrieval.

REFERENCES

Chater, N. and Vitnyi, P. M. (2003). The generalized universal law of generalization. *Journal of Mathematical Psychology*, 47:346–369.

Cormen, T. H., Leiserson, C. S., Rivest, R. L., and Stein, C. (2009). *Introduction to Algorithms*. Massachusets Institute of Technology.

Eidenberger, H. (2003). Distance measures for MPEG-7-based retrieval. In *Proceedings of the 5th ACM SIGMM International Workshop on Multimedia Information Retrieval*, MIR '03, pages 130–137, New York, NY, USA. ACM.

Eidenberger, H. (2012). *Handbook of Multimedia Information Retrieval*. atpress.

Faloutsos, C. and Lin, K.-I. (1995). Fastmap: A fast algorithm for indexing, data-mining and visualization of traditional and multimedia datasets. In *Proceedings of the 1995 ACM SIGMOD International Conference on Management of Data*, SIGMOD '95, pages 163–174, New York, NY, USA. ACM.

Milgram, J., Cheriet, M., and Sabourin, R. (2006). "One Against One" or "One Against All": Which One is Better for Handwriting Recognition with SVMs? In Lorette, G., editor, *Tenth International Workshop on Frontiers in Handwriting Recognition*, La Baule (France). Université de Rennes 1, Suvisoft. http://www.suvisoft.com.

Salton, G., Wong, A., and Yang, C. S. (1975). A vector space model for automatic indexing. *Commun. ACM*, 18(11):613–620.

Santini, S. and Jain, R. (1999). Similarity measures. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 21/9:871–883.

Shepard, R. N. (1987). Toward a universal law of generalization for psychological science. *Science*, 237/4820:1317–1323.

Tenenbaum, J. B. and Griffiths, T. L. (2001). Generalization, similarity, and bayesian interference. *Behavioral an*, 24:629–640.

Torgerson, W. S. (1952). Multidimensional scaling: I. theory and method. *Psychometrika*, 17/4:401–419.

Tversky, A. (1977). Features of similarity. *Psychological Review*, 84 (4):327–352.

Wisniewski, E. J. and Bassok, M. (1999). What makes a man similar to a tie? stimulus compatibility with comparison and integration. *Cognitive Psychology*, 39(34):208 – 238.

Yang, Y. and Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 42–49, New York, NY, USA. ACM.