

MUSE: Multi-Represented Similarity Estimation

Hans-Peter Kriegel, Peter Kunath, Alexey Pryakhin, Matthias Schubert

*Department Institute for Informatics, Ludwig-Maximilians-University Munich
Oettingenstr. 67, D-80538 Munich*

{kriegel, kunath, pryakhin, schubert}@dbs.ifi.lmu.de

Abstract—In modern multimedia databases, objects can be specified by a large variety of feature representations. In this paper, we present a novel technique for multi-represented similarity estimation. We transform the distance between two objects in each representation into so-called similarity and dissimilarity estimates which are used to derive a meaningful similarity score. To determine the parameters for our new similarity measure, we present methods with and without user feedback.

I. INTRODUCTION

To measure the similarity between two multimedia objects, modern systems often employ multiple representations, e.g. the texture and the color distribution of an image. In order to exploit multiple representations, it is necessary to select or combine the most suitable feature representations. In this paper, we introduce a new technique for estimating the similarity based on the distance values observed in multiple representations. Therefore, we introduce so-called similarity and dissimilarity estimates for each representation which mirror how strongly a distance $d_R(o_1, o_2)$ in representation R indicates similarity or dissimilarity w.r.t. an underlying notion of similarity. A key feature of this approach is that the influence of each representation is adjusted to the strength of evidence it provides for either similarity or dissimilarity. If a given distance value does not hint to either statement, it remains rather neutral and the joint similarity measure relies more on the other representations. If available, the given notion of similarity can be provided by user feedback. Additionally, we propose a method for finding a good combination rule without any training examples based on the observation that the provided representations are usually useful for the given task. Thus, two objects having a very small or a very large distance in all representations, are usually considered as similar or dissimilar. Thus, we propose an iterative process to fit both types of estimates by maximizing the agreement between the estimates in each representation. In our experimental evaluation, we demonstrate that our new method is capable to improve precision and recall of similarity queries.

II. RELATED WORK

A good overview about combining approaches in information retrieval can be found in [1]. In [2] the authors use logistic regression on user feedback to weight the influence of each representation. The authors of [3] suggest another technique calculating a weighted distance function that uses the standard deviations of the features. Further approaches to approximate weights employing user feedback are described in [4], [5].

In comparison to our approach, these methods employ global weights for each representation and no self-adjusting distance tuning is proposed.

III. SIMILARITY AND DISSIMILARITY ESTIMATES

To compare two multi-represented objects o_1 and o_2 , we have to combine the distances that can be derived from each representation. The core idea of our method is to derive two types of estimates from each representation describing the distribution of similarity and dissimilarity over the observed distance values. By combining both aspects into a joint estimate, the distance in each representation is transformed into a similarity score expressing similarity, dissimilarity or ambiguity of meaning. Afterwards, we can combine the estimates from all representations for deriving a joint similarity score.

To define our new model, we assume that there is a ground truth of the pairwise similarity between data objects o_1, o_2 which is described by a similarity score $sim_{o_1, o_2} \in [0..1]$. To define dissimilarity, we can invert sim_{o_1, o_2} by subtraction from 1, i.e. $dissim_{o_1, o_2} = 1 - sim_{o_1, o_2}$. The main problem of judging similarity is that the distance between o_1 and o_2 in representation R often mimics this ground truth only insufficiently, i.e. sim_{o_1, o_2} might vary from sim_{o_3, o_4} even if $d_R(o_1, o_2) = d_R(o_3, o_4)$.

If the majority of pairwise similarity is observed for distances $d_R(o_i, o_j) \leq \tau$, it can be concluded that the portion of similarity being observed for distances $d_R(o_i, o_k) > \tau$ is rather small. Thus, for distance values being larger than τ , the objects are likely to be rather dissimilar. Correspondingly, we can consider the portion of dissimilarity being observed for objects having a larger distance $d_R(o_i, o_j) \geq \tau$. If most of the dissimilarity corresponds to larger distance values than τ , objects having $d_R(o_1, o_2) < \tau$ will most likely be similar. The advantage of estimating similarity by the absence of dissimilarity and vice versa is that the ambiguity is captured in a natural way. An ambiguous meaning is caused by rather similar and rather dissimilar objects displaying the same distance. If a certain interval of distance values does not display a large amount of similarity, it is natural that the meaning of a distance in this interval is rather indicating object dissimilarity. To capture this idea in a mathematical formulation, we consider the total amount of similarity or dissimilarity :

$$TotSim(DB) = \sum_{o_1, o_2 \in DB} sim_{o_1, o_2}$$

Correspondingly, the total dissimilarity is defined as:

$$TotDissim(DB) = \sum_{o_1, o_2 \in DB} dissim_{o_1, o_2}$$

Now, a similarity estimate can be defined as the absence of dissimilarity and a dissimilarity estimate by the absence of similarity.

Definition 1 (Similarity and Dissimilarity Estimate): Let DB be a database with the total similarity $TotSim(DB)$ and the total dissimilarity $TotDissim(DB)$. Furthermore, let R be a representation with the distance measure $d_R : R \times R \rightarrow [0 \dots 1]$. Then, we define the set $\Phi(x) = \{(u_1, u_2) | (u_1, u_2) \in R \times R \wedge d_R(u_1, u_2) \geq x\}$ as the set of all object pairs having a distance larger than or equal to x . The similarity estimate $\varepsilon_R^S(x)$ is defined as follows:

$$\varepsilon_R^S(x) = \frac{\sum_{(o_1, o_2) \in \Phi_R(x)} dissim_{o_1, o_2}}{TotDissim(DB)}$$

Let $\Psi(x) = \{(u_1, u_2) | (u_1, u_2) \in R \times R \wedge d_R(u_1, u_2) \leq x\}$ be the set of all object pairs having a distance less than or equal to x . The dissimilarity estimate $\varepsilon_R^D(x)$ is defined as follows:

$$\varepsilon_R^D(x) = \frac{\sum_{(o_1, o_2) \in \Psi_R(x)} sim_{o_1, o_2}}{TotSim(DB)}$$

To find a joint statement about object similarity in representation R , we employ the estimate making a stronger implication because the reliability of the estimate increases with the portion of the covered similarity or dissimilarity. Furthermore, to find a joint distance measure, we invert the similarity estimate by subtracting it from one. The joint estimate $\varepsilon_R()$ for representation R is defined as follows:

Definition 2 (Joint Estimate):

Let DB be a database and let R be a representation with the distance measure $d_R : R \times R \rightarrow [0 \dots 1]$. Furthermore, let $\varepsilon_R^S(x)$ be the similarity estimate and let $\varepsilon_R^D(x)$ be the dissimilarity estimate for the distance value x . Then, the joint estimate for x is defined as follows:

$$\varepsilon_R(x) = \begin{cases} 1 - \varepsilon_R^S(x) & \text{if } \varepsilon_R^S(x) > \varepsilon_R^D(x) \\ \varepsilon_R^D(x) & \text{else} \end{cases}$$

The joint estimate represents a statement about the assumed object dissimilarity from each representation. Considering the case of missing representations as described above, the estimate distance is formulated as follows:

Definition 3 (Estimate Distance):

Let o_1 and o_2 be two multi-represented objects over the representations $\rho = \{R_1, \dots, R_n\}$. For each feature space F_i corresponding to representation R_i , we consider the joint estimate $\varepsilon_R(x)$. The estimate distance d_{EST} between o_1 and o_2 is defined as follows:

$$d_{EST} : (R_1 \times \dots \times R_n) \times (R_1 \times \dots \times R_n) \rightarrow \mathbb{R}_o^+$$

$$d_{EST}(o_1, o_2) = \left(\prod_{i \in M} \varepsilon_{R_i}(d_{R_i}(o_1, o_2)) \right)^{\frac{1}{|M|}}$$

IV. CALCULATING ESTIMATES AND PARAMETER TUNING

a) *User Feedback:* In the basic case, there exists a sufficiently large set of user feedback, i.e. there already exist pairs of distance values and similarity scores in each representation. To learn the estimates, we employ regression based on a sigmoid function which displayed the best results in our experiments.

$$sig_{\alpha, \beta}(x) = \frac{1}{1 + \exp(\alpha \cdot x + \beta)}$$

A similarity estimate is expressed by a sigmoid function having a negative α -value. Correspondingly, dissimilarity estimates are approximated by a sigmoid function having a positive α -value.

In the following, we sketch the steps to determine the parameters α and β . Given the user feedback $DB_u \subseteq DB \times DB \times [0..1]$, the following steps are performed for representation R : (1) Insert each triple $(o_1, o_2, dissim_{o_1, o_2})$ into a priority queue which is ordered w.r.t. $d_R(o_1, o_2)$ in descending order. Simultaneously, sum up $TotDissim(DB_u)$. (2) Process all elements in the priority queue. Determine the estimate points by taking the distances as x -values. The y -values are calculated by $\frac{dissim_{o_1, o_2}}{TotDissim(DB_u)}$. (3) Employ the numerical method proposed by Levenberg and Marquardt [6] for fitting the sigmoid function to the sample points.

Correspondingly, the dissimilarity estimate is calculated by ordering the priority queue in ascending order and employing similarity scores.

b) *Self Adjustment:* In the case that there is no user feedback available, we can still derive a good parameter setting by employing the following iterative algorithm. Therefore, we have to find a way to maximize the agreement between the estimates. Formally, we can capture the agreement between the estimates in all representations by the average variance of estimate values on a training data set DB :

Definition 4: Let $\rho = \{R_1, \dots, R_n\}$ be a set of representations and let $\varepsilon_{R_i}^S()$ and $\varepsilon_{R_i}^D()$ be similarity and dissimilarity estimates for all R_i with $1 \leq i \leq n$, respectively. Then the average similarity $\varepsilon_{avg}^S()$ and dissimilarity estimate $\varepsilon_{avg}^D()$ for the comparison of two multi-represented objects o_1, o_2 are defined as :

$$\varepsilon_{avg}^S(o_1, o_2) = \frac{\sum_{R_i \in R} \varepsilon_{R_i}^S(o_1, o_2)}{|R|}$$

$$\varepsilon_{avg}^D(o_1, o_2) = \frac{\sum_{R_i \in R} \varepsilon_{R_i}^D(o_1, o_2)}{|R|}$$

Consequently, the average variance of a given example set DB consisting of multi-represented objects is given by:

$$Var(DB) = \sum_{(o_n, o_m) \in DB} \sum_{R_i \in R} (\varepsilon_{R_i}^S(o_n, o_m) - \varepsilon_{avg}^S(o_n, o_m))^2 + (\varepsilon_{R_i}^D(o_n, o_m) - \varepsilon_{avg}^D(o_n, o_m))^2$$

After providing a measure of the agreement between estimates on a given example set DB , we now introduce an iterative method for finding parameters minimizing $Var(DB)$.

```

FUNCTION selfAdjustingParameterTuning(pct)
BEGIN
   $D := \text{generateDistanceVectors}()$ 
  FOR EACH representation  $R_i \in \rho$  DO
     $\text{initEstimates}(D)$ 
  END DO
  DO
     $\backslash\backslash$  built joint model
     $\varepsilon_{avg}^S := \text{calculateAVGSIM}(D)$ 
     $\varepsilon_{avg}^D := \text{calculateAVGDISSIM}(D)$ 
     $\backslash\backslash$  fit estimates to joint model
    FOR EACH representation  $R_i \in \rho$  DO
       $\text{approximateEstimates}(D, \varepsilon_{avg}^S, \varepsilon_{avg}^D)$ 
    END DO
  WHILE(old parameters  $\neq$  new parameters)
END

```

Fig. 1. The algorithm for self-adjusting parameter tuning.

Figure 1 displays our approximation algorithm. in pseudo code. The algorithm starts with an initial estimate in each representation R_i inducing ε_{avg}^S and ε_{avg}^D for all distances that can be observed in the training set. Thus, ε_{avg}^S and ε_{avg}^D are considered as the current model of the notion similarity. In a maximization step, we now adjust our estimates to resemble this notion of similarity in a best possible way. Therefore, we apply the method described in the previous subsection. However, since we do not have any definite statement about the similarity of two objects, we need to consider each object comparison for both types of estimates and assign the current estimate value as current similarity score. Thus, the comparison between o_1 and o_2 is considered to have the similarity of $\varepsilon_{avg}^S(o_1, o_2)$ when fitting the similarity estimates. Correspondingly, $\varepsilon_{avg}^D(o_1, o_2)$ is used as similarity score when fitting the dissimilarity estimates. After adjusting the estimate parameters to the current notion of similarity, we update the average estimates ε_{avg}^S and ε_{avg}^D which can be done by directly applying the definition. Both steps are continued until the estimate parameters do not change significantly any more which is usually the case after 3 to 7 iterations.

V. EVALUATION

We performed experiments on five real world data sets displaying comparable performance. However due to space restrictions, we cannot discuss all of our results. In the following, we will describe our results on one of these data sets. We compared our new approach to a weighted sum of distances (“WAVG”), the approach using logistic regression [2] (“LogReg”) and the best single representation. These methods and our new approach were implements in JAVA 1.5. The *conf-data-set* contains 183 pictures belonging to 35 classes which were Taken during two sightseeing trips. We extracted color histograms [7], gray histograms, roughness and facet-orientation [8]. We employed the iterative method without user feedback to determine the parameters for the estimate distance.

Figure 2 displays the improvement of the precision recall graph compared to WAVG, LogReg and the single representation displaying the best result. In all cases, the estimate distance showed a significant improvement of the precision of up to 4 %.

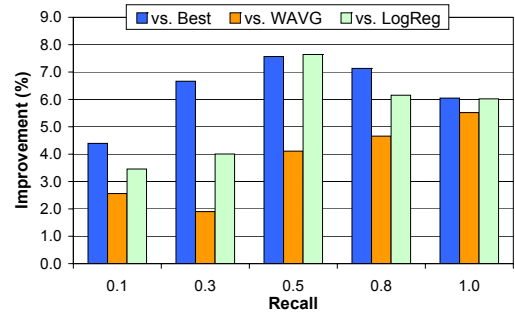


Fig. 2. Improvement of precision w.r.t. to five different levels of recall for the conf data set.

To get a further impression of the performance of our new method, we additionally employed nearest neighbor classification based on four distances. For this experiment, we employed 10-fold cross validation to test the distance measures on unknown samples as well. The estimate distance allowed distance-based classification with an accuracy of 87 %. WAVG and LogReg performed worse by showing accuracies of 84%. The best single representation outperformed the two established methods by 1% allowing an accuracy of 85 %.

VI. CONCLUSIONS

In this paper, we propose a novel technique for similarity estimation in multi-represented similarity search. Therefore, similarity and dissimilarity estimates are introduced describing the strength of the similarity and dissimilarity statements in each representation. After combining both statements for each representation, a new combined distance measure, called estimate distance, is composed. For fitting both types of estimates to a given application, we propose a method employing user feedback. Furthermore, an iterative method is proposed that does not rely on user feedback. In our experiments, the estimate distance outperforms state-of-the-art combination methods w.r.t. precision and recall.

REFERENCES

- [1] W. B. Croft, *Advances in Information Retrieval: Recent Research from the CIIR, (Chapter 1 :Combining Approaches to Information Retrieval)*. Kluwer Academic Publishers, 2000.
- [2] F. Gey, “Inferring probability of relevance using the method of logistic regression,” in *proc. 17th ACM SIGIR Conference on Research and Development in Information Retrieval, Dublin, Ireland*, 1994, pp. 222–231.
- [3] S. Aksoy, R. M. Haralick, F. A. Cheikh, and M. Gabbouj, “A weighted distance approach to relevance feedback,” in *Proc. ICPR*, 2000.
- [4] T. S. Chua, W. C. Low, and C. X. Chu, “Relevance feedback techniques for color-based image retrieval,” in *Proc. MMM*, 1998.
- [5] Y. Rui, T. S. Huang, and S. Mehrotra, “Content-based image retrieval with relevance feedback in mars,” in *Proc. ICIP*, 1997.
- [6] P. E. Gill and W. Murray, “Algorithms for the solution of the nonlinear least-squares problem,” *SIAM J. Num. Anal.*, vol. 15, no. 5, pp. 977–992, 1978.
- [7] M. Ortega, Y. Rui, K. Chakrabarti, K. Porkaew, S. Mehrotra, and T. S. Huang, “Supporting ranked boolean similarity queries in MARS,” *Knowledge and Data Engineering*, vol. 10, no. 6, pp. 905–925, 1998.
- [8] G. Chinga, O. Gregersen, and B. Dougherty, “Paper surface characterization by laser profilometry and image analysis,” *J. Mic. Anal.*, vol. 84, pp. 5–7, 2003.