Fuzzy Approach to Non-metric Similarity Indexing

Tomáš Bartoš
Charles University in Prague, Faculty of Mathematics and Physics, SIRET research group
bartos@ksi.mff.cuni.cz

Alan Eckhardt
Charles University in Prague, Faculty of Mathematics and Physics, SIRET research group
eckhardt@ksi.mff.cuni.cz

Tomáš Skopal
Charles University in Prague, Faculty of Mathematics and Physics, SIRET research group
skopal@ksi.mff.cuni.cz

ABSTRACT

The task of similarity search becomes more complex when the distance measure is not a metric. In this paper, we investigated the recently proposed *fuzzy* approach to similarity search in non-metric databases where the triangle inequality might not hold. In summary, we took nine fuzzy T-norms, proposed a tuning algorithm for the fuzzy T-norm operators (*Lambda Tuning Algorithm*), and applied this approach to the pivot-based search. We present the results focusing on the efficiency and effectiveness of the suggested method.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Retrieval Models, Search Process; H.3.1 [**Content Analysis and Indexing**]: Indexing Methods

General Terms

Theory, Experimentation, Performance

1. INTRODUCTION

Non-metric spaces are currently a big challenge in the research of similarity searching because violating at least one metric postulate makes it difficult to efficiently and effectively answer a query using a metric index. One of the most common conditions that might not hold, e.g., in multimedia retrieval, is the triangle inequality. To deal with this case, one might transform the space (distances or data) with a suitable function or use the fuzzy operators instead.

2. RELATED WORK

To the best of our knowledge, this is the first attempt to apply fuzzy concepts to similarity indexing. In [5] the authors proposed a utilization of fuzzy sets to reason about distance measures but this approach was used only to user-perceived real-world distances. Much more interest has been on the indexing in fuzzy databases (e.g. [2, 3]). However, the authors dealt with fuzzy data stored in a database, while our approach uses crisp data interpreted with fuzzy operators.

SISAP '11, June 30–July 1, 2011, Lipari, Italy.

3. INDEXING WITH FUZZY OPERATORS

In contrast to previous approaches, it was suggested to use fuzzy T-norm operators for indexing non-metric data [4]. Instead of modifying the distances or data, the "+" operator needs to be tuned, replacing the typical triangle inequality condition used in metric filtering by so called *T-norms*.

Fuzzy T-norms. A parametrized triangular norm (*T-norm*) is a binary operation $T_\lambda : [0,1]^2 \to [0,1]$, such that for a fixed λ and $\forall x, y, z \in [0,1]$ (working with similarities), the following conditions are satisfied:

(T1)	$T_\lambda(x,y) = T_\lambda(y,x)$	commutativity
(T2)	$T_\lambda(T_\lambda(x,y),z) = T_\lambda(x,T_\lambda(y,z))$	associativity
(T3)	$T_\lambda(x,y) \leq T_\lambda(x,z)$ whenever $y \leq z$	monotonicity
(T4)	$T_\lambda(x,1) = x$	boundaries

3.1 Lambda Tuning Algorithm

Finding the right λ parameter for a fuzzy T-norm operator is the crucial point when using fuzzy logic in indexing.

We propose the Lambda Tuning Algorithm (LTA), see Algorithm 1 for details, inspired by the TriGen algorithm [8]. In each iteration, it gets the error (the number of violating triplets) for the given data and the current λ value, and modifies λ parameter according to this error. While TriGen computes triplets violating the triangle inequality, we count triplets for which the triangle transitivity (defined as $sim(o,q) \geq T_\lambda(sim(o,p), sim(p,q))$) does not hold [4].

We have normalized all T-norms to accept $\lambda \in [0, \infty]$ – most flexible (min. error) for $\lambda = 0$ and most strict (max. error) for $\lambda = \infty$. The LTA returns the most suitable λ parameter for the given fuzzy T-norm operator, error threshold and given data.

Algorithm 1 LTA (T-norm t, ErrThreshold, database)

```
lMin = 0; lMax = t.getLambdaMax();
l = (lMin + lMax)/2;
bestL = −1;
for all i < MaxIter do {MaxIter is predefined and fixed}
   err = getError(t, l, data);    {violating triplets rate}
   if err <= ErrThreshold then {worsen the operator}
      bestL = lMin = l;
      l = (l + lMax)/2;
   else {improve the operator}
      lMax = l;
      l = (lMin + l)/2;
   end if
end for
return bestL; {Return the best lambda found}
```

4. EXPERIMENTS

In this section, we verify the proposed Lambda Tuning Algorithm and analyze the efficiency/effectiveness of fuzzy T-norms. As the testbed for all the experiments, we have used the CoPhIR database [1] with up to 50,000 images represented by normalized vectors with 512 dimensions. We used the fractional L_p functions[1] with $p = 0.5$ (semimetrics), and distance-to-similarity conversion $sim(q, o) = \frac{1}{1+\delta(q,o)}$.

4.1 Lambda Tuning Algorithm Results

We studied the behaviour of 9 fuzzy T-norm families (defined in [6]), namely AczelAczina, Dombi, Frank, Hamacher, MayorTorrens, Metric, SugenoWeber, SchweizerSklar, and Yager, focused mainly on upper bound limits and how they adapt to error threshold changes. The training set consisted of 500 objects and we tuned the families on 50,000 triplets for several error threshold values. The results in Figure 1 show that some families easily adapt (e.g., Yager or Metric) while others do not reflect the changes in error thresholds.

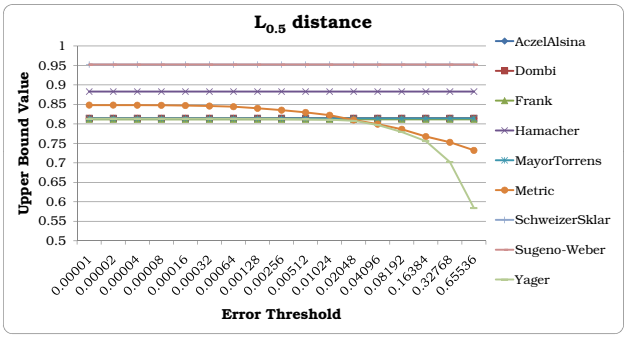

Figure 1: Upper bounds for fuzzy T-norms

4.2 Pivot Table Evaluation

We replaced the conventional Pivot table filtering [7] by fuzzy filtering, where the distance lower bound was substitued by similarity upper bound obtained by the specific T-norm. Then we tested the *efficiency* (number of distance computations, see Figure 2) and the *effectiveness* (precision of the result, see Figure 3) on range queries using 50 randomly selected objects with maximum error rate of 0.001%.

The closer look reveals that even though the fuzzy T-norms supposed to be flexible and more adaptable compared to the classic non-metric searching, the best results showed only a small speed up of query evaluation for the price of higher error rate (e.g. Hamacher, SchweizerSklar).

5. CONCLUSIONS

We investigated the possibility of applying fuzzy T-norms to indexing non-metric databases. We proposed the Lambda Tuning Algorithm to search for the best λ parameter for a fuzzy T-norm and we suggested pivot table fuzzy filtering. Unfortunately, we found out that proposed usage of T-norms is not applicable to researched data. The reason is the discrepancy between the λ parameter learnt for the given error and the (considerable) error that λ had on queries (the reason is still to be determined). Moreover, the bi-directional distance-to-similarity conversion is required for the fuzzy approach. This is a drawback for any distance measure but it might be an advantage if applying fuzzy T-norms to pure similarity functions which remains as a future work.

[1]The Minkowski distances $L_p(u, v) = \left(\sum_{i=1}^{n} |u_i - v_i|^p\right)^{1/p}$, for which the triangle inequality is violated if $p \in (0, 1)$.

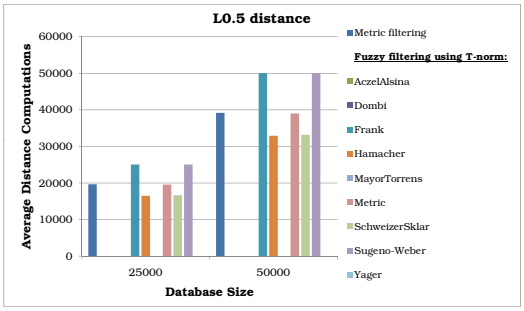

Figure 2: Average Distance Computations

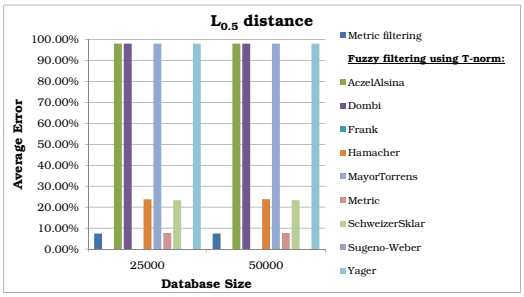

Figure 3: Average Error

6. ACKNOWLEDGMENTS

This research has been supported by Czech Science Foundation (GAČR) projects 201/09/0683 and 202/11/0968.

7. REFERENCES

[1] P. Bolettieri, A. Esuli, F. Falchi, C. Lucchese, R. Perego, T. Piccioli, and F. Rabitti. CoPhIR: a test collection for content-based image retrieval. *CoRR*, abs/0905.4627v2, 2009.

[2] P. Bosc and M. Galibourg. Indexing principles for a fuzzy data base. *Information Systems*, 14(6):493 – 499, 1989.

[3] P. Bosc, M. Galibourg, and G. Hamon. Fuzzy querying with sql: Extensions and implementation aspects. *Fuzzy Sets and Systems*, 28(3):333 – 349, 1988.

[4] A. Eckhardt, T. Skopal, and P. Vojtáš. On fuzzy vs. metric similarity search in complex databases. In *Proceedings of the 8th International Conference on Flexible Query Answering Systems*, FQAS '09, pages 64–75, Berlin, Heidelberg, 2009. Springer-Verlag.

[5] H. W. Guesgen. Reasoning about distance based on fuzzy sets. *Applied Intelligence*, 17:265–270, September 2002.

[6] E. P. Klement, R. Mesiar, and E. Pap. *Triangular Norms*. Kluwer, Dordrecht, The Netherlands, 2000.

[7] G. Navarro. Analyzing metric space indexes: What for? In *IEEE SISAP 2009*, pages 3–10, 2009.

[8] T. Skopal. Unified framework for fast exact and approximate search in dissimilarity spaces. *ACM Trans. Database Syst.*, 32, November 2007.