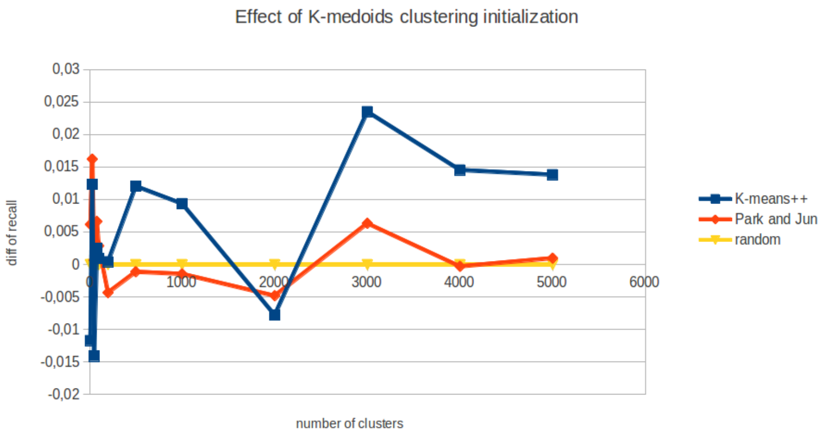

Fig. 2: Effect of the initialization procedure for the K-medoids clustering in the quality of nearest-neighbors search

solution to the problem, but it is available only for a few distance functions. In this article we propose to address that limitation, by extending LSH to general metric spaces, using K-medoids clustering as basis for a LSH family of functions. We show in our experiments that the K-medoids LSH improves the results over the random choice of sample of DFLSH, while keeping the advantage of relying only on distance information. As expected, K-medoids LSH performance is slightly worse than K-means LSH, but it is important to note that K-means relies heavily on the vector-space structure, and many data types of interest do not offer such structure.

REFERENCES

Akune, F., Valle, E., and Torres, R. MONORAIL: A Disk-Friendly Index for Huge Descriptor Databases. In *2010 20th International Conference on Pattern Recognition*. IEEE, pp. 4145–4148, 2010.

Arthur, D. and Vassilvitskii, S. k-means++: the advantages of careful seeding. In *Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. SODA '07. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, pp. 1027–1035, 2007.

Datar, M., Immorlica, N., Indyk, P., and Mirrokni, V. S. Locality-sensitive hashing scheme based on p-stable distributions. In *Proceedings of the twentieth annual symposium on Computational geometry - SCG '04*. ACM Press, New York, New York, USA, pp. 253, 2004.

Gionis, A., Indyk, P., and Motwani, R. Similarity search in high dimensions via hashing. In *Proceedings of the 25th International Conference on Very Large Data Bases*. VLDB '99. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 518–529, 1999.

Indyk, P. and Motwani, R. Approximate nearest neighbors. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing - STOC '98*. ACM Press, New York, New York, USA, pp. 604–613, 1998.

Kang, B. and Jung, K. Robust and Efficient Locality Sensitive Hashing for Nearest Neighbor Search in Large Data Sets. In *NIPS Workshop on Big Learning (BigLearn)*. Lake Tahoe, Nevada, pp. 1–8, 2012.

Kaufman, L. and Rousseeuw, P. J. *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley-Interscience, 1990.

Lowe, D. G. Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60 (2): 91–110, Nov., 2004.

Novak, D., Kyselak, M., and Zezula, P. On locality-sensitive indexing in generic metric spaces. *Proceedings of the Third International Conference on SImilarity Search and APplications - SISAP '10*, 2010.

Ostrovsky, R., Rabani, Y., Schulman, L., and Swamy, C. The Effectiveness of Lloyd-Type Methods for the k-Means Problem. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS'06)*. Vol. 59. IEEE, pp. 165–176, 2006.

Park, H.-S. and Jun, C.-H. A simple and fast algorithm for K-medoids clustering. *Expert Systems with Applications* 36 (2): 3336–3341, 2009.

Paulevé, L., Jégou, H., and Amsaleg, L. Locality sensitive hashing: A comparison of hash function types and querying mechanisms. *Pattern Recognition Letters* 31 (11): 1348–1358, Aug., 2010.

Tellez, E. S. and Chavez, E. On locality sensitive hashing in metric spaces. In *Proceedings of the Third International Conference on SImilarity Search and APplications*. SISAP '10. ACM, New York, NY, USA, pp. 67–74, 2010.