[4] H. Bisgin and H.N. Dalfes. Parallel clustering algorithms with application to climatology. In *Geophysical Research Abstracts*, volume 10, 2008.

[5] A. Broder, L. Garcia-Pueyo, V. Josifovski, S. Vassilvitskii, and S. Venkatesan. Scalable k-means by ranked retrieval. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 233–242. ACM, 2014.

[6] M.S. Charikar. Similarity estimation techniques from rounding algorithms. In *STOC*, pages 380–388, New York, NY, USA, 2002. ACM.

[7] C.L.A. Clarke, N. Craswell, I. Soboroff, and G.V. Cormack. Overview of the TREC 2010 web track. Technical report, DTIC Document, 2010.

[8] C.L.A. Clarke, N. Craswell, I. Soboroff, and E.M. Voorhees. Overview of the TREC 2011 web track. Technical report, DTIC Document, 2011.

[9] C.L.A. Clarke, N. Craswell, and E.M. Voorhees. Overview of the TREC 2012 web track. Technical report, DTIC Document, 2012.

[10] A. Coates, A. Karpathy, and A. Ng. Emergence of object-selective features in unsupervised feature learning. In *NIPS 25*, page 2690–2698, 2012.

[11] G.V. Cormack, M.D. Smucker, and C.L.A. Clarke. Efficient and effective spam filtering and re-ranking for large web datasets. *Information retrieval*, 14(5):441–465, 2011.

[12] C.M. De Vries. *Document clustering algorithms, representations and evaluation for information retrieval*. PhD thesis, Electricial Engineering and Computer Science, Queensland University of Technology, Brisbane, Australia, 2014.

[13] C.M. De Vries and S. Geva. K-tree: large scale document clustering. In *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pages 718–719. ACM, 2009.

[14] C.M. De Vries, S. Geva, and A. Trotman. Document clustering evaluation: Divergence from a random baseline. *WIR-2012, Dortmund, Germany.*, 2012.

[15] C.M. De Vries, R. Nayak, S. Kutty, S. Geva, and A. Tagarelli. Overview of the INEX 2010 XML mining track: Clustering and classification of XML documents. In *INEX 2010*, pages 363–376, 2011.

[16] R.M. Esteves and C. Rong. Using mahout for clustering wikipedia's latest articles: a comparison between k-means and fuzzy c-means in the cloud. In *IEEE CloudCom*, pages 565–569. IEEE, 2011.

[17] A. Gersho and R.M. Gray. *Vector quantization and signal compression*. Kluwer Academic Publishers, 1993.

[18] S. Geva and C.M. De Vries. Topsig: topology preserving document signatures. CIKM '11, pages 333–338, New York, NY, USA, 2011. ACM.

[19] Z.S. Harris. Distributional structure. *Word*, 1954.

[20] N. Jardine and C.J. van Rijsbergen. The use of hierarchic clustering in information retrieval. *Information storage and retrieval*, 7(5):217–240, 1971.

[21] R. Jin, C. Kou, R. Liu, and Y. Li. Efficient parallel spectral clustering algorithm design for large data sets under cloud computing environment. *Journal of Cloud Computing: Advances, Systems and Applications*, 2(1):18, 2013.

[22] W.B. Johnson and J. Lindenstrauss. Extensions of Lipschitz mappings into a Hilbert space. *Contemporary mathematics*, 26(189-206):1–1, 1984.

[23] A. Kulkarni and J. Callan. Document allocation policies for selective searching of distributed indexes. In *CIKM 2010*, pages 449–458. ACM, 2010.

[24] A. Kumar and S. Mukherjee. Verification and validation of mapreduce program model for parallel k-means algorithm on hadoop cluster. *International Journal of Computer Applications*, 72, 2013.

[25] Q.V. Le, M.A. Ranzato, R. Monga, M. Devin, K. Chen, G.S. Corrado, J. Dean, and A.Y. Ng. Building high-level features using large scale unsupervised learning. *arXiv preprint arXiv:1112.6209*, 2011.

[26] D.D. Lewis, Y. Yang, T.G. Rose, and F. Li. RCV1: A new benchmark collection for text categorization research. *The Journal of Machine Learning Research*, 5:361–397, 2004.

[27] T. Liu, C. Rosenberg, and H.A. Rowley. Clustering billions of images with large scale nearest neighbor search. In *WACV'07*, pages 28–28. IEEE, 2007.

[28] C.D. Manning, P. Raghavan, and H. Schutze. An introduction to information retrieval. 2008.

[29] R. Nayak, C.M. De Vries, S. Kutty, S. Geva, L. Denoyer, and P. Gallinari. Overview of the INEX 2009 XML mining track: Clustering and classification of XML documents. *Focused Retrieval and Evaluation*, pages 366–378, 2010.

[30] M. Sahlgren. An introduction to random indexing. In *TKE 2005*, 2005.

[31] K.C. Thompson, P Bennett, F. Diaz, C.L.A. Clarke, and E.M. Voorhees. Overview of the TREC 2013 web track. Technical report, DTIC Document, 2013.

[32] E.M. Voorhees. The philosophy of information retrieval evaluation. In *Evaluation of cross-language information retrieval systems*, pages 143–170. Springer, 2002.

[33] J. Wang, H.T. Shen, J. Song, and J. Ji. Hashing for similarity search: A survey. *arXiv preprint arXiv:1408.2927*, 2014.

[34] Xin-Jing Wang, Lei Zhang, and Ce Liu. Duplicate discovery on 2 billion internet images. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2013 IEEE Conference on*, pages 429–436. IEEE, 2013.

[35] Ren Wu, Bin Zhang, and Meichun Hsu. Clustering billions of data points using gpus. In *Proceedings of the combined workshops on UnConventional high performance computing workshop plus memory access workshop*, pages 1–6. ACM, 2009.

[36] Tian Zhang, Raghu Ramakrishnan, and Miron Livny. Birch: an efficient data clustering method for very large databases. In *ACM SIGMOD Record*, volume 25, pages 103–114. ACM, 1996.