Appendix

A. Additional similarity matrices

In Figure 10, we provide daily similarity matrices over one month. We observe that the content for weekends and weekdays might differ for both datasets. Note that the number of points per day is equalized to avoid artifacts due to the number of vectors per day.

Figure 9 presents weekly similarity matrices over one year. We observe the similar content drift behavior to Figure 4, but no discernable weekly correlations.

B. Balance of K-means clusters

The IVF-based indexing relies on a vector quantizer to partition the vectors into clusters. Therefore, we investigate how content drift affects K-means clusters. We select months i and j and train K-means (K=16384) on Φ_i. Then, we assign the vectors from Φ_j to the trained centroids, count the number of points within each cluster and normalize them by $|\Phi_j| = M$. This yields a discrete distribution $p^{i,j} = (p_1, \ldots, p_K)$ We use the entropy of $H(p^{i,j})$ to measure the balancedness of the K-means clusters. For balanced clusters the entropy is $\log_2 K = 14$ and for a hopelessly unbalanced clustering where all vectors are assigned to one cluster it is 0. Figure 8 shows the matrix of entropies for all pairs (i, j). The further away from the diagonal, the lower the entropy. This means that the K-means clustering becomes progressively less balanced when month i is more distant from month j. In addition, for YFCC, the clusters are more imbalanced for opposite seasons.

This means that the direct distance measurements in figure 4 translate to sub-optimal clustering as well. For all datasets, the content drift takes place and has different nature and behavior. The changing distribution also affects K-means clusters and hence might lead to the noticeable degradation of the most prevalent indexing schemes at scale.

C. Robustness of indexing structures for different window sizes

In our experiments, we consider the window size m=3 months which is motivated by the reasonable practical scenario. However, one can consider different m settings.

In Table 5, we provide the robustness results for IVF indexes built upon uncompressed embeddings for various window sizes m in months. We select the coarse quantizer sizes according to the number of datapoints within the index. We observe that the performance degradation does not differ much, even for large m.

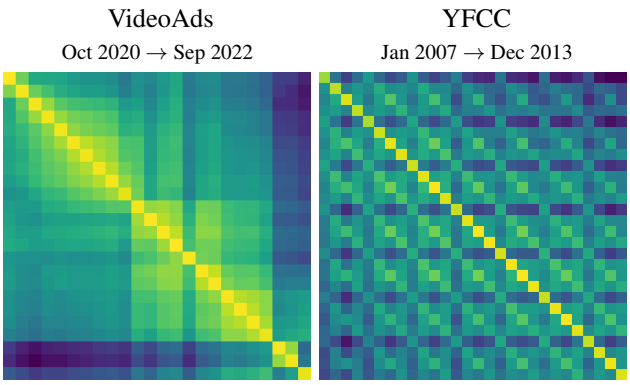

Figure 8. Balancedness of K-means clusters over time. The starting and ending date for the periods are indicated on top. For both datasets, the clusters become more imbalanced. YFCC also demonstrates the seasonal behavior — the clusters are more balanced for the same seasons than for the opposite ones. Note that we use stride 3 months for the YFCC dataset for better visualization.

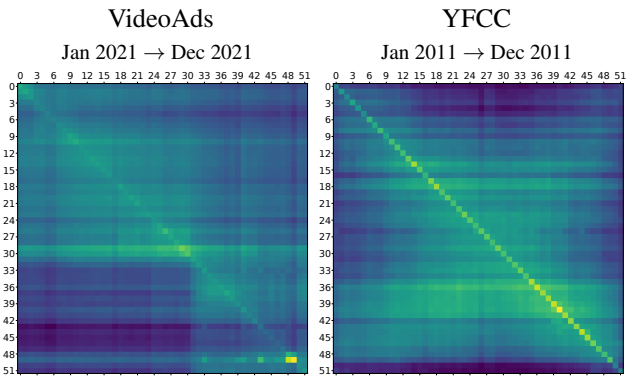

Figure 9. Pairwise similarity matrices between the embeddings over one year subdivided in one week. Blue and yellow correspond to low and high similarities, respectively. There is still the seasonal pattern for YFCC and content drift over time for VideoAds. Both datasets do not have any clearly visible weekly correlations.

D. DeDrift-Lazy with multiple training iterations

DeDrift-Lazy can be considered as a warm-started k-means to adapt to the new data distribution. Therefore, we investigate the impact of the number of centroid update steps L. For a normal k-means clustering the number of iterations strikes a tradeoff between speed and the quality of the clustering. However, Table 6 demonstrates that a single centroid update provides the highest recall. Moreover, the number of training iterations L>2 leads to noticeable degradation. This is because DeDrift-Lazy do not reassign the points after the centroid update and hence more iterations imply that the centroids move far away from the ones that the "old" vectors were assigned to. Therefore, it is both more efficient and more accurate to do a single centroid update step.

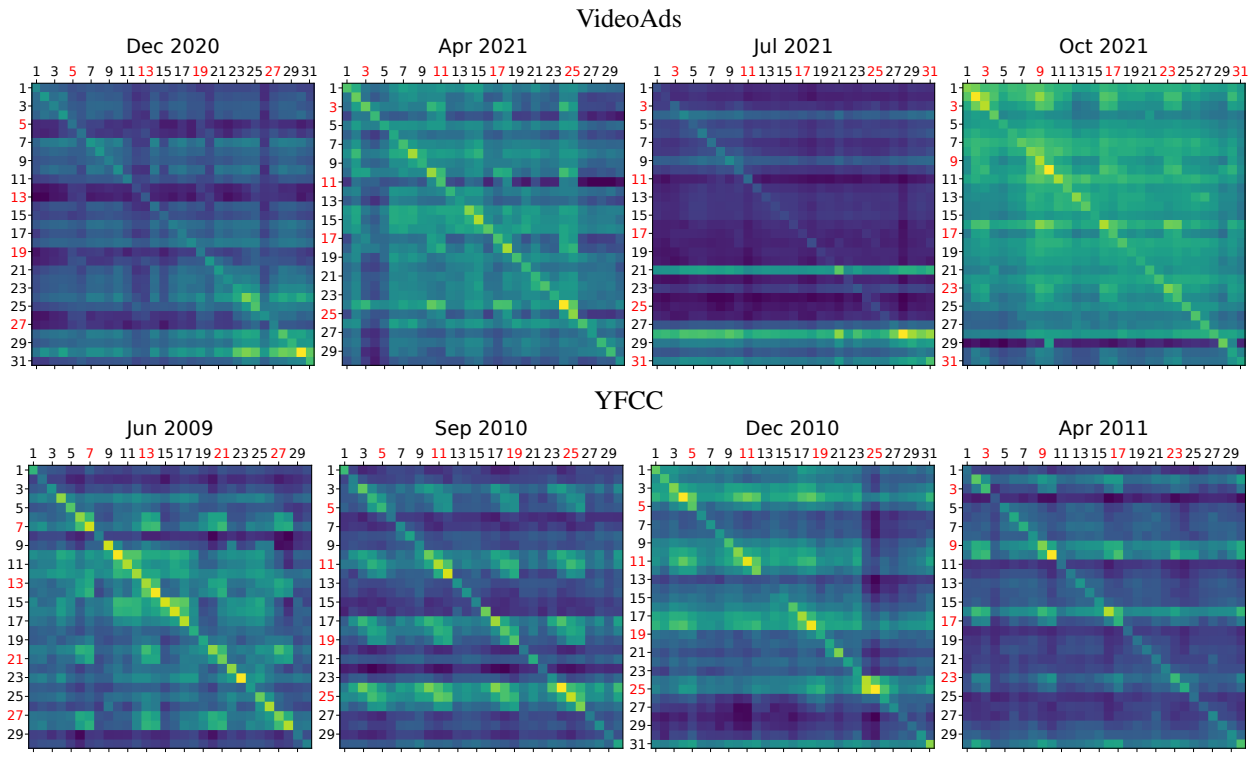

Figure 10. Pairwise similarity matrices between the embeddings over one month subdivided in days for a few months selected at random. Blue and yellow correspond to low and high similarities, respectively. Red dates represent weekends. Both datasets have noticeable weekday vs weekend pattern.

Budget		6000 DCS		12000 DCS		30000 DCS		60000 DCS	
Method	m	ID	OOD	ID	OOD	ID	OOD	ID	OOD
IVF8192	1	0.873	0.767	0.934	0.867	0.977	0.949	0.990	0.979
IVF16384	3	0.842	0.732	0.914	0.845	0.966	0.938	0.985	0.973
IVF32768	6	0.839	0.738	0.896	0.832	0.956	0.930	0.979	0.967
IVF65536	12	0.821	0.743	0.896	0.850	0.955	0.937	0.978	0.969
Budget		6000 DCS		12000 DCS		30000 DCS		60000 DCS	
Method	m	ID	OOD	ID	OOD	ID	OOD	ID	OOD
IVF2048	1	0.876	0.826	0.938	0.912	0.980	0.970	0.992	0.989
IVF4096	3	0.796	0.744	0.892	0.858	0.960	0.945	0.983	0.977
IVF8192	6	0.768	0.713	0.872	0.839	0.943	0.928	0.974	0.967
IVF16384	12	0.758	0.703	0.859	0.823	0.939	0.924	0.973	0.964

Table 5. Relative performance of IVF indexing structures for in-domain (ID) and out-of-domain (OOD) search on VideoAds (top) and YFCC (bottom) for different window sizes m in months. The search accuracy measure is 10-recall@10. The drops in performance are essentially similar for various m settings.

E. Index update costs for IVF with PQ compressed embeddings on YFCC

Table 7 provides the update costs for the IVF index with OPQ encoding. On both datasets, DEDRIFT demonstrates efficiency gains from 3× to 10×.

Note that the gains are smaller than for IVF operating on uncompressed embeddings. This is because, in this experiment, the index on the PQ compressed vectors uses original data on the disk and loads it into RAM at each update step. This is an implementation choice, that in addition makes the timings dependent on the performance of the external storage. Specifically, in our case, the data loading takes ∼1.7s and ∼8s for YFCC and VideoAds, respectively.

F. DEDRIFT on IVF with PQ compressed embeddings on YFCC

Table 8 presents the results of the IVF index with OPQ encoding on the YFCC dataset. The performance drop caused by the content drift is smaller compared to VideoAds. Nevertheless, DEDRIFT almost closes the gap between no reindexing (None) and full index reconstruction (Full).

G. Runtimes for different budgets

In this section, we report measured search times in milliseconds for different DCS budgets on each dataset. We average the runtimes over 20 independent runs. All runs are performed with 30 threads on an Intel Xeon Gold 6230R CPU @ 2.10GHz.

H. Running DEDRIFT on reconstructed vectors

In Table 9, we present the index update method performance if the cetroids are updated based on either original

YFCC, IVF4096,Flat, Jun 2013					
Budget (DCS)	6000	12000	20000	30000	60000
L=0	0.746	0.858	0.913	0.943	0.975
L=1	0.795	0.889	0.930	0.954	0.979
L=2	0.795	0.888	0.931	0.954	0.979
L=3	0.791	0.884	0.928	0.952	0.978
L=5	0.785	0.879	0.924	0.949	0.976
L=10	0.777	0.871	0.919	0.945	0.973
VideoAds, IVF16384,Flat, Jun 2022					
Budget (DCS)	6000	12000	20000	30000	60000
L=0	0.719	0.832	0.891	0.923	0.961
L=1	0.780	0.875	0.920	0.946	0.971
L=2	0.780	0.869	0.913	0.939	0.966
L=3	0.773	0.863	0.909	0.934	0.962
L=5	0.769	0.860	0.904	0.930	0.958
L=10	0.753	0.844	0.893	0.920	0.951

Table 6. DEDRIFT-Lazy performance for the different number of centroid update iterations L. L=1 provides the highest recall values. Note that L=1 is also the most efficient option.

YFCC, IVF4096,OPQ32				
Method	Split	Lazy	Hybrid	Full
Update costs (s)	2.1	8.1	10.4	29.8
VideoAds, IVF16384,OPQ32				
Method	Split	Lazy	Hybrid	Full
Update costs (s)	10.8	24.1	33.2	430.8

Table 7. Index update costs for IVF indexes with OPQ encoding. DEDRIFT variants are much more efficient than full index reconstruction (Full).

	IVF4096,OPQ32, direct encoding				
Budget (DCS)	6000	12000	20000	30000	60000
None	0.414	0.444	0.455	0.461	0.465
Split	0.423	0.448	0.457	0.461	0.465
Lazy	0.432	0.452	0.459	0.463	0.466
Hybrid	0.432	0.453	0.460	0.464	0.466
Full	0.435	0.454	0.460	0.464	0.467
	IVF4096,OPQ32, residual encoding				
Budget (DCS)	6000	12000	20000	30000	60000
None	0.453	0.481	0.492	0.497	0.501
Split	0.463	0.487	0.495	0.500	0.503
Lazy	0.474	0.495	0.504	0.507	0.510
Hybrid	0.478	0.496	0.504	0.507	0.510
Full	0.480	0.500	0.508	0.511	0.514

Table 8. Comparison of the index update methods on the YFCC dataset for IVF4096,OPQ32.

embeddings or reconstructed ones from the PQ encodings. DEDRIFT does not degrade the recall values much while the full index reconstruction is noticeably affected.

YFCC, IVF4096,OPQ32, direct encoding								
Budget	6000 DCS		12000 DCS		30000 DCS		60000 DCS	
Method	Orig	Recon	Orig	Recon	Orig	Recon	Orig	Recon
Split	0.423	0.423	0.448	0.447	0.461	0.461	0.465	0.466
Lazy	0.432	0.430	0.452	0.452	0.463	0.463	0.466	0.466
Full	0.435	0.425	0.454	0.448	0.464	0.462	0.467	0.465
VideoAds, IVF16384,OPQ32, direct encoding								
Budget	6000 DCS		12000 DCS		30000 DCS		60000 DCS	
Method	Orig	Recon	Orig	Recon	Orig	Recon	Orig	Recon
Split	0.520	0.514	0.556	0.552	0.579	0.578	0.587	0.586
Lazy	0.530	0.528	0.563	0.562	0.583	0.582	0.588	0.588
Full	0.548	0.527	0.573	0.559	0.588	0.580	0.593	0.587

Table 9. DEDRIFT and full index reconstruction performance (Full) when the centroids are updated using original embeddings (Orig) and reconstructed ones from PQ encodings (Recon).

Budget (DCS)	6000	12000	20000	30000	60000
IVF16384, Flat	6.12	12.05	18.93	27.14	53.35
IVF16384, OPQ32	1.08	1.23	1.29	1.40	1.96

Table 10. Runtimes (ms per query) for different budgets on VideoAds.

Budget (DCS)	6000	12000	20000	30000	60000
IVF4096, Flat	4.26	7.72	12.61	18.19	35.44
IVF4096, OPQ32	0.43	0.54	0.72	0.93	1.73

Table 11. Runtimes (ms per query) for different budgets on YFCC.

I. Evolving k-means evaluation

In this experiment, we evaluate evolving k-means [9] during the full index reconstruction. We consider different evolving k-means configurations proposed in the paper and provide the results in Table 12. Evolving k-means slightly improves the results on both datasets.

J. Image credits

J.1. Attributions for Figure 2

From top to bottom and left to right, the images are from Yahoo Flickr users:
2007-07: Imagine24, fsxz, Tuldas, CAPow!, Anduze traveller, Barnkat.
2008-10: BEYOND BAROQUE, armadillo444, Anadem Chung, nikoretro, Jon Delorey, Gone-Walkabout.
2009-12: Spider58, thehoneybunny, Communicore82, Oli Dunkley, HarshLight, Yelp.com.
2010-02: cruz_fr, Bemep, Dawn - Pink Chick, ljw7189, john.meagher, ShashiBellamkonda.

Budget (DCS)	6000	12000	20000	30000	60000
Full naive	0.804	0.895	0.938	0.959	0.981
Full [9] PSKV	0.798	0.892	0.935	0.958	0.982
Full [9] FSKV p=0.5	0.804	0.895	0.938	0.960	0.983
Full [9] FSKV p=0.8	**0.807**	**0.896**	**0.939**	**0.961**	**0.983**
Budget (DCS)	6000	12000	20000	30000	60000
Full naive	0.815	0.892	0.930	0.952	0.975
Full [9] PSKV	0.815	0.894	0.934	0.956	0.980
Full [9] FSKV p=0.5	0.818	0.896	0.935	0.956	0.980
Full [9] FSKV p=0.8	**0.820**	**0.897**	**0.935**	**0.957**	**0.981**

Table 12. Comparison with the evolving k-means method [9] on the YFCC (Top) and VideoAds (Bottom) datasets. Evolving k-means slightly improves the recall rates for the full index reconstruction.

2011-06: robinmyerscough, telomi, richmiller.photography, hergan family, cus73, librarywebchic.