A Case for Hubness Removal in High–Dimensional Multimedia Retrieval

Dominik Schnitzer[1], Arthur Flexer[1], and Nenad Tomašev[2]

[1] Austrian Research Institute for Artificial Intelligence (OFAI), Vienna, Austria
[2] Artificial Intelligence Laboratory, Jožef Stefan Institute, Ljubljana, Slovenia

Abstract. This work investigates the negative effects of hubness on multimedia retrieval systems. Because of a problem of measuring distances in high-dimensional spaces, hub objects are close to an exceptionally large part of the data while anti-hubs are far away from all other data points. In the case of similarity based retrieval, hub objects are retrieved over and over again while anti-hubs are nonexistent in the retrieval lists. We investigate textual, image and music data and show how re-scaling methods can avoid the problem and decisively improve the overall retrieval quality. The observations of this work suggest to make hubness analysis an integral part when building a retrieval system.

Keywords: hubness, high dimensionality, multimedia retrieval

1 Introduction

A number of publications have recently discussed hubness in a general machine learning context and introduced it as a new aspect of the curse of dimensionality [12, 14]. Hub objects are nearest neighbors to many other data points in high dimensional spaces and hence are being frequently retrieved without being semantically relevant to many of the queries. The effect is related to the phenomenon of concentration of distances and has been shown to have a negative impact on many tasks including outlier detection [20], clustering [17] and collaborative filtering [10]. More related to multimedia retrieval, the influence of hubs on object recognition [16] and music similarity [4] has also been investigated. Our work studies hubness in a range of multimedia retrieval systems (text, image, music) and shows how removal of hubness via re-scaling of the distances improves retrieval quality and overall system performance.

2 Investigation of multimedia retrieval systems

To investigate the impact of hubs on multimedia retrieval systems, we build three standard content-based multimedia retrieval systems (text, image, music). We use the systems in a query-by-example scenario where the top–k answers are retrieved and evaluated in terms of precision/recall with their class label. This is one of the most common scenarios for content-based multimedia retrieval. The three systems and eight data sets used for evaluation are described below:

— The text retrieval system analyzes textual data by first employing stopword removal, stemming and transformation to the bag-of-words representation. Standard *tf* · *idf* (term frequency · inverse document frequency) weights are computed (see e.g. [2]). The word vectors are normalized to the average document length. Document vectors are compared with the cosine distance. We evaluate our text-retrieval system with the *Twitter (C1ka)* [13] and the UCI [5] *Mini Newsgroups* data sets which both show a very high extrinsic dimensionality (Twitter: 49 000, Mini Newsgroups: 8 000).

— The image retrieval system uses standard dense SIFT vectors [8] to compute spatial histograms of visual words. Each image is represented with a bag-of-visual-word (BOVW) vector [15]. To measure the similarity between two images we compute the Euclidean distance between their BOVW vectors. We evaluate our image-retrieval system with the *Caltech 101* [3], *Leeds Butterfly* [19] and *17 Flowers* [11] data sets. The extrinsic data dimensionality of the features used in this retrieval system is 36 000.

— The music retrieval system extracts MFCC features and estimates a single multivariate Gaussian representing the timbral structure of the music piece [9]. The similarity between the Gaussian representations is computed with the Jensen-Shannon divergence. This method is one of the standard methods for content based music retrieval systems. The music-retrieval system is evaluated with the *Ballroom* [6], *GTzan* [18] and *ISMIR 2004* [1] data sets. The extrinsic data dimensionality of the features is 1 275.

2.1 The dominance of hubs in retrieval lists

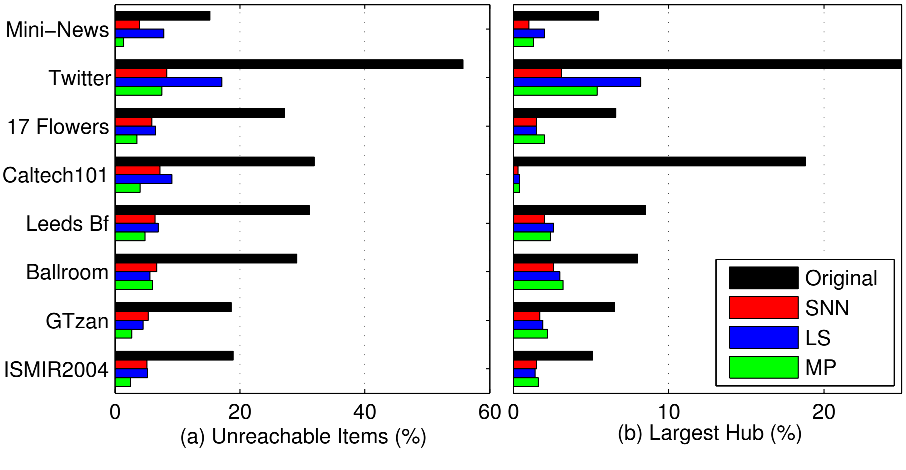

Fig. 1. At $k = 5$, (a) the number of unreachable items and (b) the size of the largest hub (as percentages of the total collection size for original and re-scaled distances).

The described three retrieval systems are based on different kinds of data (*text*, *image*, *music*), different feature representations (*vectors*, *multivariate Gaussians*) and three different distance measures (*cosine*, *Euclidean*, *Jensen-Shannon*). But all systems operate with very high dimensional features (1 275 to 49 000) which should make them prone to hubness. To verify this, we analyze the k–nearest neighbor (kNN) lists of all data objects in all data sets. We compute (for $k = 5$) the percentage of items which is unreachable in our retrieval systems (do not appear in any of the kNN lists, i.e. anti-hubs) and the maximum percentage one single item appears in all kNN lists (i.e. the largest hub). In Figure 1(a) (black bars for original distance spaces) we see that on average over 20% of all objects in our system are unreachable, i.e. will never be retrieved.

Table 1. Evaluation of multimedia retrieval systems in terms of hubness (S), precision (P), recall (R) and f–measure (F_1) for $k = 1, 5, 10$ and different distance spaces (Original, SNN, LS, MP) . Top results per line and $k = 1, 5$ or 10 in bold face.

Dataset		Original			SNN			LS			MP		
	k	1	5	10	1	5	10	1	5	10	1	5	10
Mini-	S	4.1	5.1	4.4	1.8	1.7	1.5	1.2	1.3	1.5	**0.9**	**0.6**	**0.5**
Newsgroups	P	.644	.541	.489	.597	.518	.475	.672	.571	.516	**.677**	**.576**	**.526**
text	R	.006	.027	.049	.006	.026	.047	**.007**	**.029**	.052	**.007**	**.029**	**.053**
	F_1	.013	.052	.089	.012	.049	.086	**.013**	.054	.094	**.013**	**.055**	**.096**
Twitter	S	29.2	14.6	10.7	3.9	3.7	3.2	2.3	2.9	3.1	**1.8**	**1.8**	**1.8**
(C1ka)	P	.319	.265	.245	.398	.377	.356	.478	**.473**	**.434**	**.490**	.467	**.434**
text	R	.004	.016	.028	.005	.021	.039	**.007**	**.033**	.058	**.007**	**.033**	**.059**
	F_1	.008	.029	.048	.010	.038	.067	.013	**.058**	.094	**.014**	**.058**	**.095**
17 Flowers	S	5.9	3.9	3.4	1.5	1.0	1.0	1.3	1.2	1.0	**1.0**	**0.7**	**0.5**
image	P	.520	.391	.343	.447	.367	.330	**.570**	.433	.378	.557	**.442**	**.394**
	R	.006	.024	.043	.006	.023	.041	**.007**	.027	.047	**.007**	**.028**	**.049**
	F_1	.013	.046	.076	.011	.043	.073	**.014**	.051	.084	**.014**	**.052**	**.088**
Caltech	S	80.7	50.1	34.5	1.5	1.4	1.4	1.4	1.2	1.1	**1.3**	**1.1**	**1.0**
101	P	.598	.544	.508	.591	.542	.514	**.686**	**.631**	**.587**	.672	.624	.579
image	R	.005	.021	.037	.005	.020	.037	**.007**	**.028**	**.049**	**.007**	.027	.048
	F_1	.011	.039	.064	.009	.038	.064	**.014**	**.052**	**.085**	.013	.051	.083
Leeds-	S	3.7	3.5	3.0	1.4	1.0	0.9	1.2	0.9	1.0	**1.0**	**0.5**	**0.4**
Butterfly	P	.584	.470	.409	.466	.394	.360	**.630**	**.529**	**.465**	.608	.522	.462
image	R	.007	.028	.050	.006	.024	.044	**.008**	**.032**	**.057**	.007	**.032**	.056
	F_1	.014	.054	.088	.011	.045	.078	**.015**	**.060**	**.101**	**.015**	**.060**	.100
Ballroom	S	3.4	2.8	2.5	1.3	0.9	0.9	1.3	1.2	1.1	**1.0**	**0.6**	**0.5**
music	P	.532	.457	.414	.520	.470	.439	.587	.505	.464	**.595**	**.512**	**.471**
	R	.006	.026	.047	.006	.026	.048	**.007**	.028	.052	**.007**	**.029**	**.053**
	F_1	.012	.049	.084	.011	.049	.087	**.013**	**.054**	.093	**.013**	**.054**	**.095**
GTzan	S	3.4	3.3	3.1	1.4	0.9	0.6	**0.9**	1.0	1.3	**0.9**	**0.4**	**0.4**
music	P	.774	.678	.599	.715	.640	.602	.796	.715	.636	**.804**	**.720**	**.652**
	R	**.008**	.034	.060	.007	.032	.060	**.008**	.036	.064	**.008**	**.036**	**.065**
	F_1	.015	.065	.109	.014	.061	.110	**.016**	.068	.116	**.016**	**.069**	**.119**
ISMIR	S	3.5	3.9	3.7	1.2	0.9	0.8	1.3	1.4	1.6	**1.0**	**0.7**	**0.4**
2004	P	.858	.808	.762	.824	.791	.762	**.914**	.846	.791	.903	**.847**	**.803**
music	R	.003	.015	.028	.003	.015	.028	**.004**	**.017**	.030	**.004**	**.017**	**.031**
	F_1	.007	.030	.053	.006	.029	.053	**.007**	**.032**	**.057**	**.007**	**.032**	**.057**

Figure 1(b) (black bars for original distance spaces) shows that a single object will be present in 5% (*ISMIR 2004*) to 18% (*Caltech 101*) or over 25% (*Twitter*) of all kNN lists. To quantify the amount of hubs/anti-hubs in a data set Radovanović et al. [12] defined a hubness measure (S^k). It measures the skewness of a histogram of each object's occurrence in the kNN lists. Positive S^k (skewed to the right) indicates that there is high hubness, i.e. there are many anti-hubs and few but very large hubs in the collection. Table 1, column *Original* records the measured hubness S for $k = 1, 5, 10$ and all data sets. The measured values range from 2.5 (*Ballroom*) to 80.7 (*Caltech 101*). With non-problematic hubness values ranging from 0 to 1, all our measurements indicate very high hubness explaining the impaired retrieval performance expressed in Figure 1.

2.2 Hubness removal for improved retrieval quality

In order to reduce hubness and its negative effects, two unsupervised methods to re-scale the high-dimensional distance spaces have been proposed [14]: Local Scaling (LS) and Mutual Proximity (MP). Both methods aim at repairing asymmetric nearest neighbor relations. The asymmetric relations are a direct consequence of the presence of hubs. A hub y is the nearest neighbor of x, but the nearest neighbor of the hub y is another point a ($a \neq x$). This is because hubs are by definition nearest neighbors to very many data points but only one data point can be the nearest neighbor to a hub. The principle of the scaling algorithms is to re-scale distances to enhance symmetry of nearest neighbors. A small distance between two objects should be returned only if their nearest neighbors concur. LS does that by computing a local statistic of the nearest neighbors to rescale the distances. MP assumes that all pairwise distances in a data set follow a certain distribution and computes the mutual probability that two points are true nearest neighbors. We use LS with a neighborhood range of $s = 10$ and MP with the empirical distribution. In addition to LS and MP we test if the shared-nearest neighbor approach (SNN) is also able to reduce hubness in our systems. We added SNN since there are a number of publications reporting its positive effects on high dimensional data [7] and there is apparent similarity to LS and MP. To use SNN first a neighborhood size (s) has to be set (in the following experiments we use $s = 50$). The new similarity between two points x and y is then the overlap in terms of their nearest neighbors (NN). Similarly to LS or MP, this enforces symmetric neighborhoods: $SNN(x, y) = |NN(x) \cap NN(y)|$.

In the previous section we have shown that all of the presented retrieval systems are strongly affected by hubs. Figure 1(a)(b) shows the impact of LS, MP and SNN in terms of largest hub size and the number of unreachable items (anti-hubs). Across all systems and data collections a pronounced decrease of unreachable items and hub sizes can be observed. Compared to the original system the number of unreachable items (Figure 1(a)) comes down from about 20% to ∼5%. In 7 out of 8 data sets MP yields the best results. Using any of the methods described also leads to a significant decrease of the largest hub sizes (Figure 1(b)). The most positive effect of the scaling methods can be observed in the *Caltech 101* dataset, where the size of the largest hub comes down to around

1% (from 18.8%). In 7 out of 8 cases SNN preforms best in terms of largest hub size. Table 1 also shows the results in terms of hubness (S) at different retrieval sizes ($k = 1, 5, 10$) for LS, MP and SNN. As expected from the previous results, S also decreases with all methods to much more normal levels between 1.8 and 0.4 (with the exception of SNN and LS for data set *Twitter*: 2.3 to 3.9). MP reduces the measured hubness to the lowest values.

Besides hubness, we have also evaluated the quality impact of hubness removal on the retrieval systems in terms of precision (P), recall (R) and f–measure (F_1). The results for all three systems are shown in Table 1 (best values are in bold font). Together with lower hubness, we observe increased retrieval quality across all domains when using LS and MP. In terms of F_1, MP performs best on 2 out of 2 text, 1 out of 3 image (*17 Flowers*) and 3 out of 3 music data sets, while LS is best on 2 out 3 image (*Leeds Butterfly*, *Caltech 101*) data sets. While LS and MP at the same time decrease hubness and increase the overall system performance (in terms of F_1, P and R), SNN only seems to reduce hubness and its retrieval quality often decreases even below the values achieved on the original distance spaces. For example in the case of *Mini Newsgroups*, F_1 at $k = 5$ decreases from originally 0.052 to 0.049, while it increases 0.054 with LS and to 0.055 with MP.

3 Summary

We have demonstrated the effect of hubness occurring in multimedia retrieval systems across three different domains (text, image, music) operating with high dimensional data. In the similarity based kNN retrieval systems, hubness caused about one fifth of the data to be unreachable anti-hubs which are never part of any of the retrieval lists. At the same time hub objects dominate many of the retrieval lists with hubs sometimes appearing in more than 25% of all possible lists. Application of distance re-scaling methods, most notably Local Scaling and Mutual Proximity, is able to decisively reduce hubness, increase reachability and thereby enhance the diversity of the retrieval lists. At the same time we see that by removing the hubs, the retrieval quality in terms of precision/recall across all data domains and retrieval systems is also increased. Whereas most existing retrieval systems do not take hubness into account, we hope that we made a clear case for hubness removal: many of the systems are negatively affected by hubness causing lower retrieval quality and diversity which can be avoided using simple re-scaling algorithms.

Acknowledgments. This research is supported by the Austrian Science Fund (FWF): P24095.

References

1. ISMIR 2004 Music Genre Collection. `http://ismir2004.ismir.net/genre_contest/index.htm`
2. Baeza-Yates, R., Ribeiro-Neto, B.: Modern Information Retrieval. Addison Wesley (1999)

3. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. Computer Vision and Image Understanding 106(1), 59–70 (2007)
4. Flexer, A., Schnitzer, D., Schlüter, J.: A mirex meta-analysis of hubness in audio music similarity. In: Proceedings of the 13th International Society for Music Information Retrieval Conference. pp. 175–180 (2012)
5. Frank, A., Asuncion, A.: UCI machine learning repository (2010), repository located at: `http://archive.ics.uci.edu/ml`
6. Gouyon, F., Dixon, S., Pampalk, E., Widmer, G.: Evaluating rhythmic descriptors for musical genre classification. In: Proceedings of the 25th International Conference Audio Engineering Society Conference. pp. 196–204 (2004)
7. Houle, M.E., Kriegel, H.P., Kröger, P., Schubert, E., Zimek, A.: Can Shared-Neighbor Distances Defeat the Curse of Dimensionality. In: Scientific and Statistical Database Management, Lecture Notes in Computer Science, vol. 6187, chap. 34, pp. 482–500. Springer, Berlin, Heidelberg (2010)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision 60(2), 91–110 (2004)
9. Mandel, M., Ellis, D.: Song-level features and support vector machines for music classification. In: Proceedings of the 6th International Conference on Music Information Retrieval. London, UK (2005)
10. Nanopoulos, A., Radovanović, M., Ivanović, M.: How does high dimensionality affect collaborative filtering? In: Proceedings of the third ACM conference on Recommender systems. pp. 293–296 (2009)
11. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing. pp. 722–729 (2008)
12. Radovanović, M., Nanopoulos, A., Ivanović, M.: Hubs in space: Popular nearest neighbors in high-dimensional data. Journal of Machine Learning Research 11, 2487–2531 (December 2010)
13. Schedl, M.: On the Use of Microblogging Posts for Similarity Estimation and Artist Labeling. In: Proceedings of the 11th International Society for Music Information Retrieval Conference. Utrecht, the Netherlands (August 2010)
14. Schnitzer, D., Flexer, A., Schedl, M., Widmer, G.: Local and global scaling reduce hubs in space. Journal of Machine Learning Research 13, 2871–2902 (October 2012)
15. Sivic, J., Russell, B.C., Efros, A.A., Zisserman, A., Freeman, W.T.: Discovering objects and their location in images. In: Proceedings of the Tenth IEEE International Conference on Computer Vision. vol. 1, pp. 370–377 (2005)
16. Tomašev, N., Brehar, R., Mladenić, D., Nedevschi, S.: The influence of hubness on nearest-neighbor methods in object recognition. In: Proceedings of the IEEE International Conference on Intelligent Computer Communication and Processing. pp. 367–374 (2011)
17. Tomašev, N., Radovanović, M., Mladenić, D., Ivanović, M.: The role of hubness in clustering high-dimensional data. In: Advances in Knowledge Discovery and Data Mining, pp. 183–195. Springer (2011)
18. Tzanetakis, G., Cook, P.: Musical genre classification of audio signals. IEEE Transactions on Speech and Audio Processing 10(5), 293 – 302 (July 2002)
19. Wang, J., Markert, K., Everingham, M.: Learning models for object recognition from natural language descriptions (2009)
20. Zimek, A., Schubert, E., Kriegel, H.P.: A survey on unsupervised outlier detection in high-dimensional numerical data. Statistical Analysis and Data Mining 5(5), 363–387 (2012)