Embed First, Then Predict

Shenghui Wang*, Rob Koopman**

OCLC Research, Schipholweg 99, 2316XA Leiden, The Netherlands,
*<shenghui.wang@oclc.org>, **<rob.koopman@oclc.org>

Shenghui Wang is a research scientist at the OCLC EMEA office in Leiden, Netherlands. Her research activities include text and data mining, semantic embedding as well as linked data. Shenghui earned a PhD in computer science from the University of Manchester in 2007, a master's degree in computer application technology at the University of Science and Technology of China in 2003, and a bachelor's degree in computer science at Anhui University in 2000.

Rob Koopman is an applied research architect at the OCLC EMEA office Leiden, Netherlands. His research activities include semantic embedding, data mining and enhancing data quality. He started to work at OCLC (then Pica) in 1981 and designed and developed major parts of the EMEA systems. Since 2012 he does applied data science for OCLC.

Wang, Shenghui and Rob Koopman. 2019. "Embed First, Then Predict." *Knowledge Organization* 46(5): 364-370. 13 references. DOI:10.5771/0943-7444-2019-5-364.

Abstract: Automatic subject prediction is a desirable feature for modern digital library systems, as manual indexing can no longer cope with the rapid growth of digital collections. It is also desirable to be able to identify a small set of entities (e.g., authors, citations, bibliographic records) which are most relevant to a query. This gets more difficult when the amount of data increases dramatically. Data sparsity and model scalability are the major challenges to solving this type of extreme multi-label classification problem automatically. In this paper, we propose to address this problem in two steps: we first embed different types of entities into the same semantic space, where similarity could be computed easily; second, we propose a novel non-parametric method to identify the most relevant entities in addition to direct semantic similarities. We show how effectively this approach predicts even very specialised subjects, which are associated with few documents in the training set and are more problematic for a classifier.

Received: 14 February 2019; Revised: 11 June 2019; Accepted: 20 June 2019

Keywords: entities, documents, subjects, embedding

1.0 Introduction

Because of the ever-increasing number of documents that information systems deal with, automatically identifying most relevant entities, such as authors, subjects, citations, or other documents is one of the most desirable features for many such systems. The size of search space is normally enormous. For example, many knowledge organisation systems (e.g., thesauri, subject heading systems) that are used in digital libraries to describe the subjects of the bibliographic records often contain tens or hundreds of thousands of terms. The number of authors or citations in a medium-to-large scale bibliographic collection reaches hundreds of thousands easily. Automatically identifying a small set of highly relevant entities from such huge search spaces—the Extreme Multi-label Text Classification (XMTC) problem—is, therefore, very difficult. Data sparsity and scalability are the major challenges.

In this paper, we describe our two-step approach to addressing this problem. First, we propose a novel embedding method which embeds different types of entities including documents themselves in the same semantic space. This method extends random projection by projecting raw entity embeddings orthogonally to an average vector, thus improving the discriminating power of resulting entity embeddings, and build more meaningful document embeddings by assigning appropriate weights to individual entities. Secondly, we propose a novel non-parametric method to predict more relevant entities for unseen documents in addition to leveraging direct semantic similarities. We compare this method with the state-of-the-art deep learning method and the direct entity-document-similarity based method.

2.0 Related work

Our goal is to automatically identify a small subset of highly relevant entities from tens or hundreds of thousands of candidates. This remains a difficult problem and is a form of Extreme Multi-label Text Classification (XMTC)

(Prabhu and Varma 2014, Bhatia et al. 2015; Liu et al. 2017), where the prediction space normally consists of hundreds of thousands to millions of labels and data sparsity and scalability are the major challenges. Different from traditional binary or multi-class classification problems, this problem of extreme multi-label text classification cannot assume that the target labels are independent or mutually exclusive. Scalable solutions became available only in recent years (Bhatia et al. 2015, Prabhu and Varma 2014). There are four categories of solutions: 1) 1-vs-all (Prabhu et al. 2018); 2) embedding-based (Bhatia et al. 2015); 3) tree-based (Prabhu and Varma 2014); and, 4) deep learning methods (Joulin et al. 2016; Liu et al. 2017). However, the performance on large-scale datasets remains low according to Bhatia el al. (2019).

3.0 Method

In our study, there are two categories of entities: 1) simple entities, such as terms (words or phrases), authors, subjects or citations; and, 2) composite entities, such as documents or bibliographic records that simple entities are associated with. Our task is to predict the most relevant simple entities to a composite entity. We propose to embed these two categories of entities in a single semantic space. This allows us to use semantic similarity to identify the most relevant simple entities to a composite query document. In addition, we propose a non-parametric prediction method that computes similarities between the query document and previously seen documents to better assess the relevance of an entity to the query document.

3.1 Ariadne semantic embedding

Let a document be a set of words and entities for which co-occurrence is relevant. In general, a document could, therefore, be a sentence, a paragraph, a fixed-size window, or, in our case, a composite bibliographic record. Let n_E be the total number of "frequent" simple entities—which could be terms (words or phrases), subjects, authors, citations—we want to embed, and D the chosen dimensionality of the embedding vectors. An entity is considered frequent when it occurs in more than K documents in the corpus, where K is flexible depending on the size of the corpus.

Building on the previous work (Koopman et al. 2015, 2017, 2019) we embed the relevant entities by Random Projection (Achlioptas 2003, Johnson and Lindenstrauss 1984) of their weighted co-occurrences, as shown in Figure 1.

Here, C is the co-occurrence matrix of different types of simple entities, R is a random matrix and C' is the matrix of final embedding vectors. Traditional random projection starts by computing the co-occurrence matrix C of size $n_E \times n_E$. This matrix contains, for each pair of entities the number of documents (or paragraphs, or sentences) of the corpus in which both entities occur. Using a matrix of random projection vectors R of size $n_E \times D$,we can then project our n_E dimensional representation of each entity to a lower D dimensional space. By leveraging the linear nature of the matrix multiplication, we can update C' directly as we go through the corpus, without ever explicitly representing C. Koopman et al. (2019) has

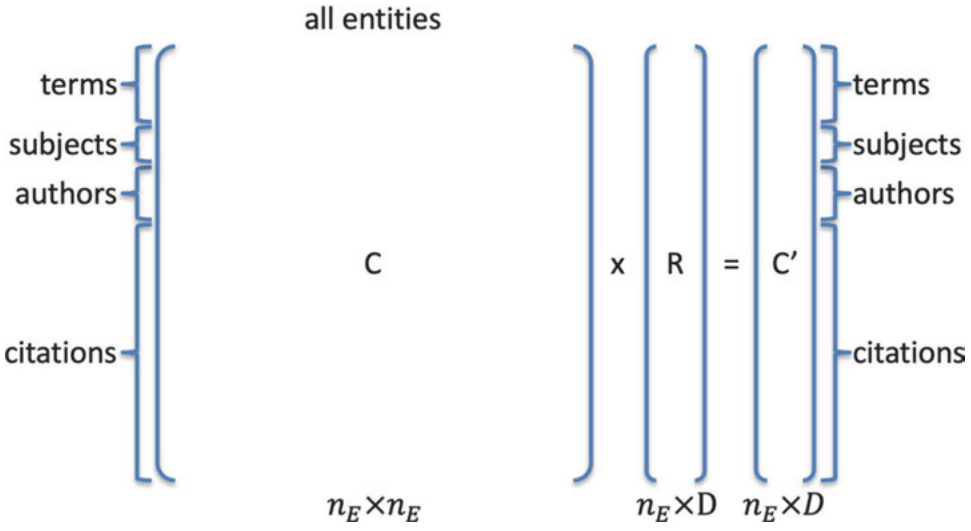

Figure 1. Random projection.

shown that this method is simple but highly efficient and scalable compared to other complicated methods while the competitive results are achieved at a fraction of the computational cost.

3.2 Orthogonal projection

Traditional models discard both very infrequent words (because they are too rare for the model to be able to capture their semantics from the training data) and very frequent words (so-called "stop words" because they do not provide any semantically useful information). In our approach, we give a continuous weight to entities based on how frequently they occur and compute the "average vector" of the corpus, $\overrightarrow{v_a}$, the sum of all the rows of C'. Unsurprisingly, this vector is very similar to the average vector of stop words. Intuitively, entities are increasingly more informative as they differ more from the average vector. By this reasoning, we project entity vectors $\overrightarrow{v_e}$ on the orthogonal hyperplane to $\overrightarrow{v_a}$:

$$\overrightarrow{v_e^*} = \overrightarrow{v_e} - (\overrightarrow{v_e} \cdot \overrightarrow{v_a})\, \overrightarrow{v_a}\,,$$

resulting in a representation where the uninformative component of entities is eliminated and normalise the vectors to have unit length. When computing document vectors, we down-weight entities according to their similarity to $\overrightarrow{v_a}$. This step is crucial to get distinctive document embeddings.

3.3 Weight assignment

Using the projection described above, the component that differentiates an entity from the average vector is kept as its final embedding. Similarly, how different an entity is from $\overrightarrow{v_a}$ also indicates how much that it contributes to the semantics of a document it is part of. In fact, we can interpret the cosine similarity as a lower bound on the mutual information (MI) between the two vectors (Foster and Grassberger 2011). In order to give a higher weight to the most informative entities, we assign a higher weight to entities with lower MI to $\overrightarrow{v_a}$ by setting the final weight of each term to be: $w_e = 1 - \cos(\overrightarrow{v_e}, \overrightarrow{v_a})$.

3.4 Document embedding

With the embeddings of the frequent simple entities and their proper weights, we can compute document embedding as the weighted average of its component entities' embeddings. Note, entities and document vectors all have unit length, making similarity computations elegant and effective.

3.5 Prediction by entity-document similarity

Once simple entities and documents are all embedded in the same semantic space, it is straightforward to calculate the similarity between any simple entity and any document. Our naïve assumption is that such similarity reflects the relevance of an entity to a document.

3.6 NPP: non-parametric prediction

We now propose a non-parametric algorithm for prediction. The algorithm returns a ranked list of entities, where the entities are sorted according to a summation of: 1) the similarity of each entity to the document; and, 2) the similarity of those of the k most similar documents from the training set which are associated with the entity. This combination provides us with a robust ranking measure, which combines the direct embedding of the entity in the semantic space where the documents also live and an extra component which lets the k nearest neighbour documents of the new document vouch for the validity of the entity. The idea is that the embedding of each document is more precise than the embedding of the individual entities (since that is done based on a combination of many documents), making the similarity computation more trustworthy and the entities those documents are associated with reflect more likely to fit the target document.

4.0 Dataset and experiments

The ASTRO dataset (available via http://www.topic-challenge.info/) contains bibliographic information of 111,616 articles published between 2003-2010 in fifty-nine astronomy and astrophysics journals indexed by the Web of Science and assigned by *Journal Citation Reports* to the astronomy and astrophysics subject field. This data set was split into the training set (containing 102,869 articles) and the testing set (containing 5,455 articles). In the training set, each article has a title, an abstract, a journal ISSN, in average 7.6 authors, 39.5 citations and 10.1 subjects. Before embedding, infrequent entities that occur in less than ten articles were discarded. Table 1 lists some stats about these entities.

Entity	#Total	#Frequent	#Frequent per record
Subject	93,566	10,624	8.7
Author	87,637	17,765	6.7
Citation	891,827	88,510	23.9
Term	105,062	25,292	45.0

Table 1. ASTRO dataset stats.

Different types of entities including terms extracted from the titles and abstracts, authors, citations and subjects were all embedded as 256-dimensional vectors using "random projection" based on their co-occurrences. The embedding of each entity was projected orthogonal to the average vector and its weight was assigned based on its similarity to the average vector as described above. When calculating the embedding for each article in the training set, we computed the weighted average of all the component simple entities. Now different types of entities and articles were embedded in the same semantic space, where their similarity could be computed easily.

Each unseen article in the testing set was also embedded into the same semantic space, as the weighted average of all the component simple entities except those that need to be predicted. For example, when predicting authors, we compute the embedding of an article based its subjects, citations and the terms in its title and abstract.

We then used entity-document similarity and the NPP algorithm to predict the most relevant entities to the articles in the testing dataset. We also applied fastText (Joulin et al. 2016) which is a state-of-the-art multi-label text classifier to the training set and used the trained model to predict the most relevant entities for the articles in the testing set. We compared the predictions from these three methods.

All the experiments were carried out on the same server with two Intel Xeon Silver 4109T 8-core processors and 384GB memory. The training process took fastText more than forty minutes to finish, while it only took our embedding method thirteen seconds to embed both simple (terms, authors, citations, subjects) and composite entities (articles). This further demonstrates the high efficiency of our embedding method as reported by Koopman et al. (2019).

5.0 Evaluation

Our task is to provide a shortlist of potentially relevant entities to the document at hand. It is important to present a ranked shortlist of candidate entities and to evaluate the quality of the prediction with an emphasis on the relevance of the top portion of such lists. Therefore, we use rank-based evaluation metrics such as precision and recall at top n. Precision@n is the proportion of the predicted entities in the top n list that are actual entities of the test document, while Recall@n is the proportion of the correctly predicted entities over all actual entities of the test document.

Figure 2 shows the Precision@n and Recall@n of three methods, where *Ariadne* represents the straightforward predictions based on entity-document similarities and Ariadne+NPP represents the non-parametric algorithm on top of Ariadne embeddings.

We can see the similar patterns across three types of predictions. Both precision and recall of the entities predicted by Ariadne is higher than those generated by fastText. The clear winner is however the Ariadne+NPP method. The precision and recall are both significantly higher than those of the other two methods, especially in terms of recall. For subject prediction, the Recall@100 is 22% higher than FastText and 11% higher than Ariadne. The advantage over FastText is more prominent for citation and author prediction (37% and 27% higher in terms of Recall@100, respectively).

For subject prediction, the Recall@100 of the Ariadne+NPP method reaches 81.3%, which is much higher than those of author and citation prediction which are slightly above 50%. In terms of Precision@20, although the absolute value of 30%, the citation prediction is nearly 10% and 20% better than subject and author prediction, respectively.

5.1 A closer look

Let us look more carefully at one concrete example article (Willis et al. 2010):

> Title: The International DORIS Service (IDS): Toward maturity
> Abstract: DORIS is one of the four space-geodetic techniques participating in the Global Geodetic Observing System (GGOS), particularly to maintain and disseminate the Terrestrial Reference Frame as determined by International Earth rotation and Reference frame Service (IERS). A few years ago, under the umbrella of the International Association of Geodesy, a DORIS International Service (IDS) was created in order to foster international cooperation and to provide new scientific products. This paper addresses the organizational aspects of the IDS and presents some recent DORIS scientific results. It is for the first time that, in preparation of the ITRF2008, seven Analysis Centers (AC's) contributed to derive long-term time series of DORIS stations positions. These solutions were then combined into a homogeneous time series IDS-2 for which a precision of less than 10 mm was obtained. Orbit comparisons between the various AC's showed an excellent agreement in the radial component, both for the SPOT satellites (e.g. 0.5–2.1 cm RMS for SPOT-2) and Envisat (0.9–2.1 cm RMS), using different software packages, models, corrections and analysis strategies. There is now a wide international participation within IDS that should lead to future improvements in DORIS analysis strategies and DORIS-derived geodetic products.

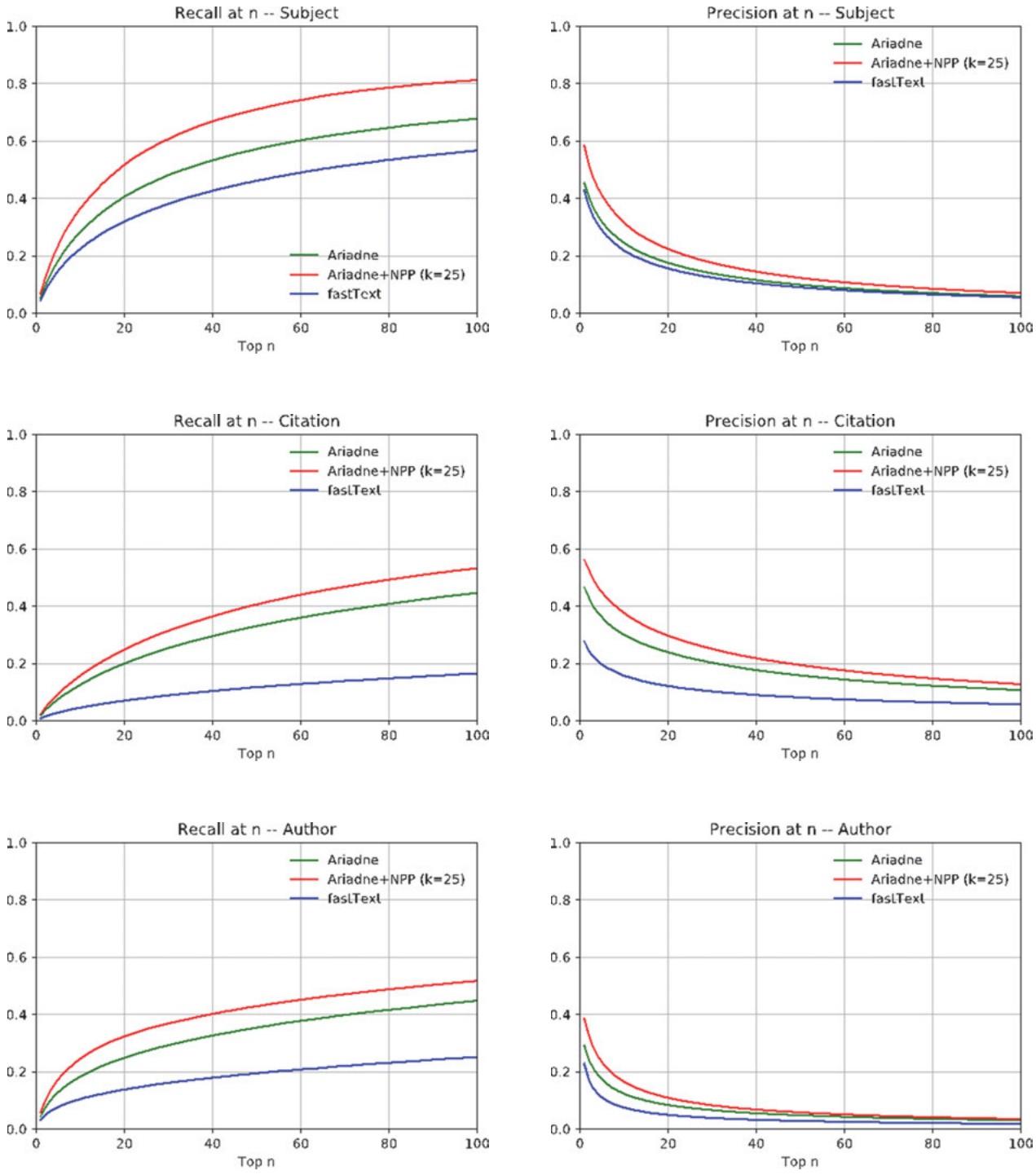

Figure 2. Precision/Recall @ n for author, citation and subject prediction.

Table 2 lists its actual subjects and the predictions by the three methods. The first column gives the raw document counts of the actual subjects in the training set. Half of the actual subjects occurred in less than thirty articles in the training set, some of which only occurred in a handful of articles or never occurred before. These extremely infrequent subjects are difficult to predict in general. FastText tends to predict common subjects, such as "model" correctly, but "earth" and "system" incorrectly (see the document counts in the last column). Even if predicted correctly, these common subjects are less informative about the article itself. Ariadne successfully predicts more specific infrequent subjects, such as "doris" and "terrestrial reference frame" but misses common ones such as

d_t	Actual subjects	Ariadne	Ariadne + NPP (k=25)	d_t	FastText	d_t
27	doris	**doris**	**doris**	**25**	gps	94
1	geodetic applications	**terrestrial reference frame**	**topex/poseidon**	**22**	earth	587
0	global geodetic observing system	**topex/poseidon**	**terrestrial reference frame**	**12**	system	1717
35	information	service	**orbit determination**	**67**	reference systems	121
3	itrf2005	geodesy	service	**23**	**doris**	25
664	mission	**precise orbit determination**	**model**	**4739**	astrometry	778
4739	model	polar motion	system	**1717**	gaia	28
204	network	**orbit determination**	**precise orbit determination**	15	**model**	4739
67	orbit determination	thermospheric model	geodesy	**13**	precession	128
15	precise orbit determination	grace	**network**	204	orbit	188
129	pressure	envisat	gps	94	**topex/poseidon**	22
302	satellite	sea level	polar motion	**11**	methods:data analysis	1734
12	terrestrial reference frame	gps	**pressure**	**129**	radiation belts	25
22	topex/poseidon	tracking	sea level	19	space	1024
		champ	envisat	14	service	23

Table 2. Comparison between fourteen actual subjects versus the top fifteen predicted ones by Ariadne, Ariadne + NPP ($k = 25$), and FastText, where the ones in bold match the actual subjects. The raw document counts of the actual subjects in the training set and those predicted by Ariadne + NPP and FastText are also given.

"model." Our Ariadne+NPP method manages to predict more specific but infrequently subjects as well as the common ones too. This makes the Recall@15 as 50%, and Precision@15 as 46.7%.

We realise that this evaluation has its limitations. As shown in Table 2, highly related subjects such as "geodesy," "gps," "polar motion" and "envisat" (Environmental Satellite) are predicted as good candidates for this article. These subjects are reasonable and potentially useful, but since they are not the subjects that the authors and human indexers have chosen, their value cannot be easily assessed. This illustrates how precision/recall may not be a very meaningful evaluation metric in this application.

That being said, we believe our predictions are still useful in practice when the predicted subjects are presented to authors or human indexers as candidate subjects to choose from. A high recall is more important as it would greatly reduce the search space and also provide opportunities for the authors and human indexers to find more suitable subjects that they probably have not thought of themselves. We believe this is also the case for author and citation prediction. In the future, we will involve subject specialists and domain experts to conduct such qualitative evaluations.

6.0 Conclusion

In this paper, we proposed a two-step approach to addressing the problem of identifying most relevant entities to a query document. We have shown that a similarity-based method based on a suitable semantic space that allows for the embedding of different types of entities is very competitive with the state-of-the-art multi-label classifier. We have described such an embedding and have shown how effective this specific semantic space really is. In addition, we proposed a novel, non-parametric, similarity-based method with the documents instead of the individual entities. We have shown that this method substantially improves the quality of the predictions, both in comparison to the state-of-the-art and to the bare similarity-based method. We also showed how our non-parametric method is particularly effective at correctly predicting very specialised subjects, which are associated with few documents in the training set and are more problematic for a classifier.

In the future, we will evaluate our method using the multi-label datasets available from the Extreme Classification Repository (Bahtia et al. 2019) and conduct more human-involved qualitative evaluation.

References

Bhatia, Kush, Himanshu Jain, Purushottam Kar, Manik Varma and Prateek Jain. 2015. "Sparse Local Embeddings for Extreme Multilabel Classification." In *Proceedings of the 28th International Conference on Neural Information Processing Systems*, ed. C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama and R. Garnett. Cambridge, MA: MIT Press, 730-38.

Dimitris Achlioptas. 2003. "Database-friendly Random Projections: Johnson-Lindenstrauss with Binary Coins." *Journal of Computer and System Sciences* 66, no. 4: 671-87.

Foster, David V. and Peter Grassberger. 2011. "Lower Bounds on Mutual Information." *Physical Review* E 83, 010101(R).

Johnson, William B. and Joram Lindenstrauss. 1984. "Extensions of Lipschitz Mappings into a Hilbert Space." *Contemporary Math* 26: 189-206.

Joulin, Armand, Edouard Grave, Piotr Bojanowski and Tomas Mikolov. 2016. "Bag of Tricks for Efficient Text Classification". In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics. Volume 2, Short Papers*, ed. Mirella Lapata, Phil Blunsom and Alexander Koller. Stroudsburg, PA: Association for Computational Linguistics, 427-31.

Koopman, Rob, Shenghui Wang, Andrea Scharnhorst, and Gwenn Englebienne. 2015. "Ariadne's Thread: Interactive Navigation in a World of Networked Information." In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*. New York: ACM, 1833-8.

Koopman, Rob, Shenghui Wang and Andrea Scharnhorst. 2017. "Contextualization of Topics: Browsing Through the Universe of Bibliographic Information." *Scientometrics* 111:1119-39.

Koopman, Rob, Shenghui Wang and Gwenn Englebienne. 2019. "Fast and Discriminative Semantic Embedding." In *Proceedings of the 13th International Conference on Computational Semantics*, Long Papers, ed. Simon Dobnik, Stergios Chatzikyriakidis and Vera Demberg. Gothenburg: Association for Computational Linguistics, 235-46.

Kush Bhatia, Kunal Dahiya, Himanshu Jain, Yashoteja Prabhu and Manik Varma. 2019. "The Extreme Classification Repository: Multi-label Datasets & Code," accessed Jan. 30. http://manikvarma.org/downloads/XC/XMLRepository.html.

Liu, Jingzhou, Wei-Cheng Chang, Yuexin Wu and Yiming Yang. 2017. "Deep Learning for Extreme Multi-label Text Classification." In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM, 115-24.

Prabhu, Yashoteja, Anil Kag, Shrutendra Harsola, Rahul Agrawal and Manik Varma. 2018. "Parabel: Partitioned Label Trees for Extreme Classification with Application to Dynamic Search Advertising." In *Proceedings of the 2018 International World Wide Web Conference*. Geneva, Switzerland: International World Wide Web Conferences Steering Committee, 993-1002. DOI: 10.1145/3178876.3185998

Prabhu, Yashoteja and Manik Varma. 2014. "FastXML: A Fast, Accurate and Stable Tree-classifier for eXtreme Multi-label Learning." In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. New York: ACM, 263-72. DOI: 10.1145/2623330.2623651

Willis, Pascal, Hervé Fagard, Pascale Ferrage, Frank G. Lemoine, Carey E. Noll, Ron Noomen, Michiel Otten, John C. Riesi, Markus Rothacher, Laurent Soudarin, Gilles Tavernierd and Jean-JacquesValette. 2010. "The International DORIS Service (IDS): Toward Maturity." *Advances in Space Research* 45:1408-20.