Random Indexing Revisited

Behrang QasemiZadeh*

National University of Ireland, Galway

University of Passau, Germany

behrang.qasemizadeh@insight-centre.org

Abstract. Random indexing is a method for constructing vector spaces at a reduced dimensionality. Previously, the method has been proposed using Kanerva's sparse distributed memory model. Although intuitively plausible, this description fails to provide mathematical justification for setting the method's parameters. The random indexing method is revisited using the principles of sparse random projections in Euclidean spaces in order to complement its previous delineation.

Keywords: Random Indexing; Dimensionality Reduction Techniques, Random Projections, Vector Space Models, Text Analytics.

1 Introduction

In order to model any aspect of language, data-driven approaches to natural language processing exploit patterns of co-occurrences. For example, distributional semantic models collect patterns of co-occurrences and investigate similarities in these patterns to quantify meanings. Vector spaces are mathematically well-defined models that are often employed to serve this purpose [18].

In a vector space model (VSM), each element $\vec{s_i}$ of its standard basis—informally, each dimension of the VSM—represents a contextual element. Given n context elements, linguistic entities are expressed using vectors $\vec{v}$ as linear combinations of $\vec{s_i}$

* This publication has emanated from research conducted with the financial support of Science Foundation Ireland under Grant Number SFI/12/RC/2289.

and scalars $\alpha_i \in \mathbb{R}$ such that $\vec{v} = \alpha_1\vec{s}_1 + \cdots + \alpha_n\vec{s}_n$. The value of α_i is acquired from the frequency of the co-occurrences of the entity that $\vec{v}$ represents and the context element that $\vec{s}_i$ represents. Therefore, the values assigned to the coordinates of a vector—that is, α_i—exhibit the correlation of an entity and context elements in an n-dimensional real vector space $\mathbb{R}^n$. In this VSM, a distance function, therefore, is employed for the discovery of similarities. Amongst several choices of distance metrics, the Euclidean distance is an innate choice. A VSM is endowed with the ℓ_2 norm to estimate distances between vectors, which is accordingly called a Euclidean VSM (denoted by $\mathbb{E}^n$). A classic document-by-term model is, perhaps, the most familiar example of the models described above for constructing VSMs [17].

In distributional approaches to text analysis, when the number of entities in a VSM increases, the number of context elements employed for capturing similarities between them surges. As a result, high-dimensional vectors, in which most elements are zero, represent entities. But, the proportional impact of context elements on similarities declines when their number increases. In a high-dimensional model, except vectors vary in most dimensions, it becomes difficult to distinguish similarities [2]. Moreover, the high-dimensionality of vectors hampers the computation of distances. These setbacks are known as the *curse of dimensionality*. A *dimensionality reduction* technique is often employed to solve these problems.

Dimensionality reduction can be achieved using a number of methods as an auxiliary process followed by the construction of a VSM. This process improves the computational performance by reducing the number of context elements employed for the construction of a VSM. In its simple form, dimension reduction can be performed by choosing a subset of context elements using a heuristic-based *selection process*. That is, a number of context elements that account for the most discriminative information in VSM are chosen using a heuristic such as a statistical weight threshold. Alternatively, a *transformation* method can be employed. This process maps $\mathbb{R}^n$ onto a $\mathbb{R}^m$, $m \ll n$, in which $\mathbb{R}^m$ is the best approximation of $\mathbb{R}^n$ in a *sense*. For example, the well-known latent semantic analysis method

employs singular value decomposition (SVD) truncation, in which $\mathbb{R}^m$ gives the best approximation of the Euclidean distances in $\mathbb{R}^n$ [7].

The use of these dimension reduction methods is hindered by a number of factors. Firstly, a VSM at the original high dimension must be first constructed. The VSM's dimension is then reduced in an independent process. Hence, the VSM at a reduced dimensionality is available for processing only after the whole sequence of these processes. Construction of the VSM at its original dimension is computationally expensive and a delay in access to the VSM at the reduced dimension is not desirable.

Secondly, reducing the dimension of vectors using the methods listed above is resource intensive. For instance, SVD truncation demands a process of the time complexity $\mathrm{O}(n^2m)$ and space complexity $\mathrm{O}(n^2)$. Similarly, depending on the employed heuristic, a selection process can be resource intensive too. Last but not least, these methods are *data-sensitive*: if the structure of the data being analysed changes—that is, if either the entities or context elements are updated—the dimensionality reduction process is required to be repeated and reapplied to the whole VSM in order to reflect the updates. As a result, these methods may not be desirable in several applications, particularly when dealing with frequently-updated big text-data.

Random projections (RPs) are employed to implement alternative dimensionality reduction methods. In the remaining of this paper, I describe the use of RPs in Euclidean spaces, which consequently arrives to the well-known random indexing (RI) technique, which has been employed in a number of applications (e.g., [3,5,19]). I then suggest a guideline for setting the method's parameters.

2 Random Projections in Euclidean Spaces

In Euclidean spaces, RPs are elucidated using the Johnson and Lindenstrauss lemma (JL lemma) [9]. Given an ϵ, $0 < \epsilon < 1$, the JL lemma states that for any set of p vectors in an $\mathbb{E}^n$, there exists a mapping onto an $\mathbb{E}^m$, for $m \geq m_0 =$

$O(^{\log p}/_{\epsilon^2})$, that does not distort the distances between any pair of vectors, with high probability, by a factor more than $1 \pm \epsilon$. This mapping is given by

$$\mathbf{M}'_{p\times m} = \mathbf{M}_{p\times n}\mathbf{R}_{n\times m}, m \ll p, n, \tag{1}$$

where $\mathbf{R}_{n\times m}$ is called the RP matrix, and $\mathbf{M}_{p\times n}$ and $\mathbf{M}'_{p\times m}$ denote the p vectors in $\mathbb{E}^n$ and $\mathbb{E}^m$, respectively. According to the JL lemma, if the distance between any pair of vectors $\vec{v}$ and $\vec{u}$ in $\mathbf{M}$ is given by the $d_{\text{Euc}}(\vec{v}, \vec{u})$, and their distance in $\mathbf{M}'$ is given by $d'_{\text{Euc}}(\mathbf{v}, \mathbf{u})$, then there exists an $\mathbf{R}$ such that $(1-\epsilon)d'_{\text{Euc}}(\mathbf{v}, \mathbf{u}) \leq d_{\text{Euc}}(\mathbf{v}, \mathbf{u}) \leq (1+\epsilon)d'_{\text{Euc}}(\mathbf{v}, \mathbf{u})$. Accordingly, instead of the original high-dimensional $\mathbb{E}^n$ and at the expense of negligible amount of error ϵ, the distance between $\vec{v}$ and $\vec{u}$ can be calculated in $\mathbb{E}^m$ to reduce the computational cost of processes.

The JL lemma does not specify the projection matrix $\mathbf{R}$. Establishing a random matrix $\mathbf{R}$ is therefore the most important design decision when using RPs. In [9], the lemma was proved using an orthogonal projection. Subsequent studies simplified the original proof that resulted in projection techniques with enhanced computational efficiency (see [4] for references). Recently, it is shown that a sparse $\mathbf{R}$, whose elements r_{ij} are defined as

$$r_{ij} = \sqrt{s}\begin{cases} -1 & \text{with probability } \frac{1}{2s} \\ 0 & \text{with probability } 1-\frac{1}{s} \\ 1 & \text{with probability } \frac{1}{2s} \end{cases}, \tag{2}$$

for $s \in \{1, 3\}$, results in a mapping that also satisfies the JL lemma [1]. Subsequent research showed that $\mathbf{R}$ can be constructed from even sparser vectors than what is suggested in [1]. In [12], it is proved that in a mapping of an n-dimensional real vector space by a sparse $\mathbf{R}$, the JL lemma holds as long as $s = O(n)$, such as $s = \sqrt{n}$ or even $s = {}^{n}/_{\log(n)}$. The sparseness of $\mathbf{R}$ consequently enhances the time and space complexity of the method by the factor $\frac{1}{s}$.

Another benefit when computing $\mathbf{M}'$ is obtained using the linearity of matrix multiplication. As stated earlier, each vector $\vec{v}_{e_i}$ in $\mathbb{E}^n$ (i.e., the ith row of $\mathbf{M}$) is given by a linear combination of the basis vectors $\vec{v}_{e_i} = w_{i1}\vec{s}_{c_1} + \cdots + w_{in}\vec{s}_{c_n}$ ($i \leq p$ and $j \leq n$). By the basic properties of the matrix multiplication, the projection of $\vec{v}_{e_i}$ in $\mathbf{M}'$ is given by $\vec{v}'_{e_i} = \vec{v}_{e_i}\mathbf{R} = w_{i1}\vec{s}_{c_1}\mathbf{R} + \cdots + w_{in}\vec{s}_{c_n}\mathbf{R}$. In turn, since by

definition all the elements of $\vec{s}_{c_k}$ are zero except the kth element (i.e., 1), $\vec{v}'_{e_i}$ can be equally written as

$$\vec{v}'_{e_i} = w_{i1}\vec{r}_1 + \cdots + w_{in}\vec{r}_n, \tag{3}$$

where $\vec{r}_j$ is the jth row of $\mathbf{R}$. Equation 3 means that row vectors $\mathbf{v}'_{e_i}$, thus $\mathbf{M}'$, can be computed directly without necessarily constructing the whole matrix $\mathbf{M}$. The jth row of $\mathbf{R}_{n\times m}$ represents a context element in the original VSM that is located at the jth column of $\mathbf{M}_{p\times n}$. Therefore, an entity at a reduced dimension can be computed directly by accumulating the row vectors of $\mathbf{R}$ that represent the context elements that co-occur with the entity.

The explanations above results in a two-step procedure similar to what is earlier suggested as the RI technique [10][16]: the construction of (a) *index vectors* and (b) *context vectors*. In the first step, each context element is assigned *exactly* to one *index vector*. [16] indicates that index vectors are high-dimensional randomly generated vectors, in which most of the elements are set to 0 and only *a few* to 1 and -1. In the second step, the construction of *context vectors*, each target entity is assigned to a vector of which all elements are zero and has the same dimension as the index vectors. For each occurrence of an entity (represented by $\vec{v}_{e_i}$) and a context element (represented by $\vec{r}_{c_k}$), the context vector is accumulated by the index vector (i.e., $\vec{v}_{e_i} = \vec{v}_{e_i} + \vec{r}_{c_k}$). The result is a vector space model constructed directly at reduced dimension. As can be understood, the first step of RI is equivalent to the construction of the random projection matrix $\mathbf{R}$, whose elements are given by Equation 2. Each index vector is a row of the random projection matrix $\mathbf{R}$. The second step of RI deals with the computation of $\mathbf{M}'$. Each context vector is a row of $\mathbf{M}'$, which is computed by the iterative process justified in Equation 3.

Compared to the justification of RI, which are based on Kanerva's sparse distributed memory (e.g., [10,16]), and whereas in previous research the method's parameters are left to be decided through experiments (e.g., [13,14]), we leverage the adopted mathematical framework to provide a guideline for setting these pa-

rameters. In an RI-constructed VSM at reduced dimension m (i.e., $\mathbb{E}^m$), the degree of preservation of distances in $\mathbb{E}^n$ and $\mathbb{E}^m$ is determined by the number of vectors in the model and the value of m. If the number of vectors is fixed, then the larger m is, the better the Euclidean distances are preserved at the reduced dimension m. In other words, the probability of preserving the pairwise distances increases as m increases. Hence, m can be seen as the capacity of an RI-constructed VSM for accommodating new entities. Compared to $m = 4000$ suggested in [10] or $m = 1800$ in [16], depending on the number of entities that are modelled in an experiment, m can be set to a smaller value such as 400.

Based on the proofs in [12], when embedding $\mathbb{E}^n$ onto $\mathbb{E}^m$, the JL lemma holds as long as s in Equation 2 is $\mathrm{O}(n)$. In text processing applications, the number of context elements (i.e., n) is often very large. When using RI, therefore, even a careful choice such as $s = \sqrt{n}$ in Equation 2 results in highly-sparse index vectors. Hence, by setting only 2 or 4 non-zero elements in index vectors, distances in the RI-constructed $\mathbb{E}^m$ resembles distances in $\mathbb{E}^n$. If the dimension of index vectors (i.e., m) is fixed, then increasing the number of non-zero elements in index vectors causes additional distortions in pairwise distances. For index vectors of fixed dimensionality m, if the number of non-zero elements increases, then the probability of the orthogonality between index vectors decreases; hence, it stimulates distortions in pairwise distances (although in some applications, distortions in pairwise distances can be beneficial).

Lastly, it is important to note that RI-constructed VSMs can be only used for estimating similarity measures that are derived from the ℓ_2 norm. For instance, the use of RI-constructed VSMs for estimating city block distances (e.g., as suggested in [11]) is not justified, at least mathematically. Hence, techniques other than RI must be used (e.g., see [8,6,21,20]).[1]

[1] An extension to this discussion and some empirical experiments can be seen in [15].

References

1. Achlioptas, D.: Database-friendly random projections. In: Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems. pp. 274–281. PODS '01, ACM, New York, NY, USA (2001)
2. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Proceedings of the 7th International Conference on Database Theory. pp. 217–235. ICDT '99, Springer-Verlag, London, UK, UK (1999), http://dl.acm.org/citation.cfm?id=645503.656271
3. Damljanovic, D., Petrak, J., Lupu, M., Cunningham, H., Carlsson, M., Engstrom, G., Andersson, B.: Random indexing for finding similar nodes within large rdf graphs. In: Proceedings of the 8th International Conference on The Semantic Web. pp. 156–171. ESWC'11, Springer-Verlag, Berlin, Heidelberg (2012), http://dx.doi.org/10.1007/978-3-642-25953-1_13
4. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Structures & Algorithms 22(1), 60–65 (2003)
5. De Vries, C.M., De Vine, L., Geva, S.: Random indexing k-tree. CoRR abs/1001.0833 (2010), http://arxiv.org/abs/1001.0833
6. De Vries, C.M., Geva, S.: Pairwise similarity of TopSig document signatures. In: Proceedings of the Seventeenth Australasian Document Computing Symposium. pp. 128–134. ADCS '12, ACM, New York, NY, USA (2012)
7. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by Latent Semantic Analysis. Journal of the American Society of Information Science 41(6), 391–407 (1990), http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.49.7546
8. Geva, S., De Vries, C.M.: TOPSIG: Topology preserving document signatures. In: Proceedings of the 20th ACM International Conference on Information and Knowledge Management. pp. 333–338. CIKM '11, ACM, New York, NY, USA (2011)
9. Johnson, W., Lindenstrauss, J.: Extensions of Lipschitz mappings into a Hilbert space. In: Conference in modern analysis and probability (New Haven, Conn., 1982), Contemporary Mathematics, vol. 26, pp. 189–206. American Mathematical Society (1984), http://www.ams.org/books/conm/026/
10. Kanerva, P., Kristoferson, J., Holst, A.: Random indexing of text samples for Latent Semantic Analysis. In: Proceedings of the 22nd Annual Conference of the Cognitive Science Society. pp. 103–6. Erlbaum (2000), http://www.rni.org/kanerva/cogsci2k-poster.txt
11. Lapesa, G., Evert, S.: Evaluating neighbor rank and distance measures as predictors of semantic priming. In: Proceedings of the Fourth Annual Workshop on Cognitive Modeling and Computational Linguistics (CMCL). pp. 66–74. Association for Computational Linguistics, Sofia, Bulgaria (August 2013), http://www.aclweb.org/anthology/W13-2608
12. Li, P., Hastie, T.J., Church, K.W.: Very sparse random projections. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 287–296. KDD '06, ACM, New York, NY, USA (2006)

13. Lupu, M.: On the usability of random indexing in patent retrieval. In: Hernandez, N., Jäschke, R., Croitoru, M. (eds.) Graph-Based Representation and Reasoning, pp. 202–216. LNCS, Springer International Publishing (2014)
14. Polajnar, T., Clark, S.: Improving distributional semantic vectors through context selection and normalisation. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014). ACL, Gothenburg, Sweden (2014), http://www.cl.cam.ac.uk/%7Esc609/pubs/eacl14tam.pdf
15. QasemiZadeh, B.: Random indexing explained with high probability. In: TSD'15. Springer (2015)
16. Sahlgren, M.: An introduction to random indexing. In: Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005 (2005), http://soda.swedish-ict.se/221/1/RI_intro.pdf
17. Salton, G., Wong, A., Yang, C.S.: A vector space model for automatic indexing. Commun. ACM 18(11), 613–620 (Nov 1975)
18. Turney, P.D., Pantel, P.: From frequency to meaning: vector space models of semantics. J. Artif. Int. Res. 37(1), 141–188 (Jan 2010), http://dl.acm.org/citation.cfm?id=1861751.1861756
19. Zadeh, B.Q., Handschuh, S.: Evaluation of technology term recognition with random indexing. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (eds.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). European Language Resources Association (ELRA), Reykjavik, Iceland (May 2014), http://www.lrec-conf.org/proceedings/lrec2014/pdf/920_Paper.pdf
20. Zadeh, B.Q., Handschuh, S.: Random Manhattan indexing. In: 25th International Workshop on Database and Expert Systems Applications. pp. 203–208. DEXA'14, IEEE (2014), http://dx.doi.org/10.1109/DEXA.2014.51
21. Zadeh, B.Q., Handschuh, S.: Random Manhattan integer indexing: Incremental L1 normed vector space construction. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1713–1723. Association for Computational Linguistics (2014), http://aclweb.org/anthology/D14-1178