| **Title**       | Random Manhattan Integer Indexing: Incremental L1 Normed Vector Space Construction |
|---------------|----------------------------------------------------------------------------------|
| **Author(s)** | QasemiZadeh, Behrang; Handschuh, Siegfried                                        |
| **Publication Date** | 2014                                                                              |
| **Publication Information** | QasemiZadeh, Behrang; Handschuh, Siegfried (2014) Random Manhattan Integer Indexing: Incremental L1 Normed Vector Space Construction Empirical Methods in Natural Language Processing (EMNLP) Doha, Qatar, 2014-10-25- 2014-10-29 |
| **Link to publisher's version** | http://emnlp2014.org/papers/pdf/EMNLP2014178.pdf |
| **Item record** | http://hdl.handle.net/10379/4650                                                  |

Some rights reserved. For more information, please see the item record link above.
Abstract

Vector space models (VSMs) are mathematically well-defined frameworks that have been widely used in the distributional approaches to semantics. In VSMs, high-dimensional vectors represent linguistic entities. In an application, the similarity of vectors—and thus the entities that they represent—is computed by a distance formula. The high dimensionality of vectors, however, is a barrier to the performance of methods that employ VSMs. Consequently, a dimensionality reduction technique is employed to alleviate this problem. This paper introduces a novel technique called Random Manhattan Indexing (RMI) for the construction of \( \ell_1 \) normed VSMs at reduced dimensionality. RMI combines the construction of a VSM and dimension reduction into an incremental and thus scalable two-step procedure. In order to attain its goal, RMI employs the sparse Cauchy random projections. We further introduce Random Manhattan Integer Indexing (RMII): a computationally enhanced version of RMI. As shown in the reported experiments, RMI and RMII can be used reliably to estimate the \( \ell_1 \) distances between vectors in a vector space of low dimensionality.

1 Introduction

Distributional semantics embraces a set of methods that decipher the meaning of linguistic entities using their usages in large corpora (Lenci, 2008). In these methods, the distributional properties of linguistic entities in various contexts, which are collected from their observations in corpora, are compared to quantify their meaning. Vector spaces are intuitive, mathematically well-defined frameworks to represent and process such information.\(^1\) In a vector space model (VSM), linguistic entities are represented by vectors and a distance formula is employed to measure their distributional similarities (Turney and Pantel, 2010).

In a VSM, each element \( \vec{s}_i \) of the standard basis of the vector space (informally, each dimension of the VSM) represents a context element. Given \( n \) context elements, an entity whose meaning is being analyzed is expressed by a vector \( \vec{v} \) as a linear combination of \( \vec{s}_i \) 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 \( \alpha_i \) is derived from the frequency of the occurrences of the entity that \( \vec{v} \) represents in/with the context element that \( \vec{s}_i \) represents. As a result, the values assigned to the coordinates of a vector (i.e. \( \alpha_i \)) exhibit the correlation of entities and context elements in an \( n \)-dimensional real vector space \( \mathbb{R}^n \). Each vector can be written as a \( 1 \times n \) row matrix, e.g. \((\alpha_1, \cdots, \alpha_n)\). Therefore, a group of \( m \) vectors in a vector space is often represented by a matrix \( M_{m \times n} \).

Latent semantic analysis (LSA) is a familiar technique that employs a word-by-document VSM (Deerwester et al., 1990).\(^2\) In this word-by-document model, the meaning of words (i.e. the linguistic entities) is described by their occurrences in documents (i.e. the context elements). Given \( m \) words and \( n \) distinct documents, each word is represented by an \( n \)-dimensional vector \( \vec{v}_i = (\alpha_{i1}, \cdots, \alpha_{in}) \), where \( \alpha_{ij} \) is a numeric value that associates the word \( \vec{v}_i \) represents to the document \( d_j \), for \( 1 < j < n \). For instance, the value of \( \alpha_{ij} \) may correspond to the frequency of the word in the document. It is hypothesized that the relevance of words can be assessed by counting the documents in which they co-occur. Therefore, words with similar vectors are assumed to have the same meaning (Figure 1).

---

\(^1\)Amongst other representation frameworks.
\(^2\)See Martin and Berry (2007) for an overview of the mathematical foundation of LSA.
In order to assess the similarity between vectors, a vector space $V$ is endowed with a norm structure. A norm $\|\cdot\|$ is a function that maps vectors from $V$ to the set of non-negative real numbers, i.e., $V \rightarrow [0, \infty)$. The pair of $(V, \|\cdot\|)$ is then called a normed space. In a normed space, the similarity between vectors is assessed by their distances. The distance between vectors is defined by a function that satisfies certain axioms and assigns a real value to each pair of vectors, i.e.

$$
\text{dist} : V \times V \rightarrow \mathbb{R}, \quad d(\vec{v}, \vec{u}) = \|\vec{v} - \vec{u}\|. \quad (1)
$$

The smaller the distance between two vectors, the more similar they are.

Euclidean space is the most familiar example of a normed space. It is a vector space that is endowed by the $\ell_2$ norm. In Euclidean space, the $\ell_2$ norm—which is also called the Euclidean norm—of a vector $\vec{v} = (v_1, \cdots, v_n)$ is defined as

$$
\|\vec{v}\|_2 = \sqrt{\sum_{i=1}^{n} v_i^2}. \quad (2)
$$

Using the definition of distance given in Equation 1 and the $\ell_2$ norm, the Euclidean distance is measured as

$$
dist_2(\vec{v}, \vec{u}) = \|\vec{v} - \vec{u}\|_2 = \sqrt{\sum_{i=1}^{n} (v_i - u_i)^2}. \quad (3)
$$

In Figure 1, the dashed line shows the Euclidean distance between the two vectors. In $\ell_2$ normed vector spaces, various similarity metrics are defined using different normalization of the Euclidean distance between vectors, e.g. the cosine similarity.

The similarity between vectors, however, can also be computed in $\ell_1$ normed spaces. The $\ell_1$ norm for $\vec{v}$ is given by

$$
\|\vec{v}\|_1 = \sum_{i=1}^{n} |v_i|, \quad (4)
$$

where $|\cdot|$ signifies the modulus. The distance in an $\ell_1$ normed vector space is often called the Manhattan or the city block distance. According to the definition given in Equation 1, the Manhattan distance between two vectors $\vec{v}$ and $\vec{u}$ is given by

$$
dist_1(\vec{v}, \vec{u}) = \|\vec{v} - \vec{u}\|_1 = \sum_{k=1}^{n} |v_k - u_k|. \quad (5)
$$

In Figure 1, the collection of the dash-dotted lines is the $\ell_1$ distance between the two vectors. Similar to the $\ell_2$ spaces, various normalizations of the $\ell_1$ distance define a family of $\ell_1$ normed similarity metrics.

As the number of text units that are being modelled in a VSM increases, the number of context elements that are required to be utilized to capture their meaning escalates. This phenomenon is explained using power-law distributions of text units in context elements (e.g. the familiar Zipfian distribution of words). As a result, extremely high-dimensional vectors, which are also sparse—i.e. most of the elements of the vectors are zero—represent text units. The high dimensionality of the vectors results in setbacks, which are colloquially known as the curse of dimensionality. For instance, in a word-by-document model that consists of a large number of documents, a word appears only in a few documents, and the rest of the documents are irrelevant to the meaning of the word. Few common documents between words results in sparsity of the vectors; and the presence of irrelevant documents introduces noise.

Dimension reduction, which usually follows the construction of a VSM, alleviates the problems

---

3The definition of the norm is generalized to $\ell_p$ spaces with $\|\vec{v}\|_p = (\sum_i |v_i|^p)^{1/p}$, which is beyond the scope of this paper.

4As long as the axioms in the distance definition hold.
listed above by reducing the number of context el-
ements that are employed for the construction of
the VSM. In its simple form, dimensionality re-
cuction can be performed using a selection pro-
cess: choose a subset of contexts and eliminate
the rest using a heuristic. Alternatively, transfor-
mation methods can be employed. A transforma-
tion method maps a vector space \( V_n \) onto a \( V_m \) of
lowered dimension, i.e. \( \tau : V_n \mapsto V_m, m \ll n. \)
The vector space at reduced dimension, i.e. \( V_m \),
is often the best approximation of the original \( V_n \)
in a sense. LSA employs a dimension reduction
technique called truncated singular value decom-
position (SVD). In a standard truncated SVD, the
transformation guarantees the least distortion in
the \( \ell_2 \) distances.\(^5\)

Besides the problem of high computational
complexity of SVD computation,\(^6\) which can be
addressed by incremental techniques (see e.g.
Brand (2006)), matrix factorization methods such
as truncated SVD are data-sensitive: if the struc-
ture of the data being analyzed changes, i.e. when
either the linguistic entities or context elements
are updated, e.g. some are removed or new ones
are added, the transformation should be recom-
puted and reapplied to the whole VSM to reflect
the updates. In addition, a VSM at the original
high dimension must be first constructed. Follow-
ning the construction of the VSM, the dimension
of the VSM is reduced in an independent process.
Therefore, the VSM at reduced dimension is avail-
able for processing only after the whole sequence
of these processes. Construction of the VSM at
its original dimension is computationally expen-
sive and a delay in access to the VSM at reduced
dimension is not desirable. Hence, the application
of truncated SVD is not suitable in several appli-
cations, particularly when dealing with frequently
updated big text–data such as applications in the
web context.

Random indexing (RI) is an alternative method
that solves the problems stated above by combing
the construction of a vector space and the di-
imensionality reduction process. RI, which is in-
roduced in Kanerva et al. (2000), constructs a
VSM directly at reduced dimension. Unlike meth-
ods that first construct a VSM at its original high
dimension and conduct a dimensionality reduction

\(^5\)Please note that there are matrix factorization techniques
that guarantee the least distortion in the \( \ell_1 \) distances, see e.g.
Kwak (2008).

\(^6\)Matrix factorization techniques, in general.

afterwards, the RI method avoids the construction
of the original high-dimensional VSM. Instead, it
merges the vector space construction and the di-
imensionality reduction process. RI, thus, signifi-
cantly enhances the computational complexity of
deriving a VSM from text. However, the applica-
tion of the RI technique (likewise the standard
truncated SVD in LSA) is limited to \( \ell_2 \) normed
spaces, i.e. when similarities are assessed using a
measure based on the \( \ell_2 \) distance. It can be verified
that using RI causes large distortions in the \( \ell_1 \)
distances between vectors (Brinkman and Charikar,
2005). Hence, if the similarities are computed us-
ing the \( \ell_1 \) distance, then the RI technique is not
suitable for the VSM construction.

Depending on the distribution of vectors in a
VSM, the performance of similarity measures
based on the \( \ell_1 \) and the \( \ell_2 \) norms varies from one
to another. For instance, it is known that the
\( \ell_1 \) distance is more robust to the presence of
outliers and non-Gaussian noise than the \( \ell_2 \) dis-
tance (e.g. see the problem description in Ke and
Kanade (2003)). Hence, the \( \ell_1 \) distance can be
more reliable than the \( \ell_2 \) distance in certain appli-
cations. For instance, Weeds et al. (2005) suggest
that the \( \ell_1 \) distance outperforms other similarity
metrics in a term classification task. In another
experiment, Lee (1999) observed that the \( \ell_1 \) dis-
tance gives more desirable results than the Cosine
and the \( \ell_2 \) measures.

In this paper, we introduce a novel method
called Random Manhattan Indexing (RMI). RMI
constructs a vector space model directly at re-
duced dimension while it preserves the pairwise
\( \ell_1 \) distances between vectors in the original high-
dimensional VSM. We then introduced a computa-
tionally enhanced version of RMI called Ran-
dom Manhattan Integer Indexing (RMII). RMI
and RMII, similar to RI, merge the construction of
a VSM and dimension reduction into an incre-
mental and thus efficient and scalable process.

In Section 2, we explain and evaluate the RMI
method. In Section 3, the RMII method is ex-
plained. We compare the proposed method with
RI in Section 4. We conclude in Section 5.

2 Random Manhattan Indexing

We propose the RMI method: a novel technique
that adapts an incremental procedure for the con-
struction of \( \ell_1 \) normed vector spaces at a reduced
dimension. The RMI method employs a two-step

procedure: (a) the creation of index vectors and (b) the construction of context vectors.

In the first step, each context element is assigned exactly to one index vector \( \vec{r}_i \). Index vectors are high-dimensional and generated randomly such that entries \( r_{ij} \) of index vectors have the following distribution:

\[
    r_{ij} = \begin{cases} 
    \frac{1}{\pi i} & \text{with probability } \frac{1}{2} \\
    0 & \text{with probability } 1 - s \\
    \frac{1}{\sqrt{2}} & \text{with probability } \frac{1}{2}
    \end{cases}, \tag{6}
\]

where \( U_1 \) and \( U_2 \) are independent uniform random variables in \((0, 1)\). In the second step, each target linguistic entity that is being analyzed is assigned to a context vector \( \vec{v}_c \) in which all the elements are initially set to 0. For each encountered occurrence of a linguistic entity and a context element—e.g. through a sequential scan of an input text collection—\( \vec{v}_c \) that represents the linguistic entity is accumulated by the index vector \( \vec{r}_i \) that represents the context element, i.e. \( \vec{v}_c = \vec{v}_c + \vec{r}_i \). This process results in a VSM of a reduced dimensionality that can be used to estimate the \( \ell_1 \) distances between linguistic entities. In the constructed VSM by RMI, the \( \ell_1 \) distance between vectors is given by the sample median (Indyk, 2000). For given vectors \( \vec{u} \) and \( \vec{v} \), the approximate \( \ell_1 \) distance between vectors is estimated by

\[
    \hat{L}_1(\vec{u}, \vec{v}) = \text{median}\{|u_i - v_i|, i = 1, \ldots, m\}, \tag{7}
\]

where \( m \) is the dimension of the VSM constructed by RMI, and \(|.|\) denotes the modulus.

RMI is based on the random projection (RP) technique for dimensionality reduction. In RP, a high-dimensional vector space is mapped onto a random subspace of lowered dimension expecting that—with a high probability—relative distances between vectors are approximately preserved. Using the matrix notation, this projection is given by

\[
    M'_m = M'_p \times R_n, \quad m \ll p, n, \tag{8}
\]

where \( R \) is often called the random projection matrix, and \( M \) and \( M' \) denote \( p \) vectors in the original \( n \)-dimensional and reduced \( m \)-dimensional vector spaces, respectively.

In RMI, the stated mapping in Equation 8 is given by Cauchy random projections. Indyk (2000) suggests that vectors in a high-dimensional space \( \mathbb{R}^n \) can be mapped onto a vector space of lowered dimension \( \mathbb{R}^m \) while the relative pairwise \( \ell_1 \) distances between vectors are preserved with a high probability. In Indyk (2000, Theorem 3) and Indyk (2006, Theorem 5), it is shown that for an \( m \geq m_0 = \log(1/\delta)^O(1/\epsilon) \), where \( \delta > 0 \) and \( \epsilon \leq 1/2 \), there exists a mapping from \( \mathbb{R}^n \) onto \( \mathbb{R}^m \) that guarantees the \( \ell_1 \) distances between any pair of vectors \( \vec{u} \) and \( \vec{v} \) in \( \mathbb{R}^n \) after the mapping does not increase by a factor more than \( 1 + \epsilon \) with constant probability \( \delta \), and it does not decrease by more than \( 1 - \epsilon \) with probability \( 1 - \delta \).

In Indyk (2000), this projection is proved to be obtained using a random projection matrix \( R \) that has Cauchy distribution—i.e. for \( r_{ij} \) in \( R \), \( r_{ij} \sim C(0, 1) \). Since \( R \) has a Cauchy distribution, for every two vectors \( \vec{u} \) and \( \vec{v} \) in the high-dimensional space \( \mathbb{R}^n \), the projected differences \( x = \vec{u} - \vec{v} \) also have Cauchy distribution, with the scale parameter being the \( \ell_1 \) distances, i.e. \( x \sim C(0, \sum_{i=1}^{n}|u_i - v_i|) \). As a result, in Cauchy random projections, estimating the \( \ell_1 \) distances boils down to the estimation of the Cauchy scale parameter from independent and identically distributed (i.i.d.) samples \( x \). Because the expectation value of \( x \) is infinite,\(^7\) the sample mean cannot be employed to estimate the Cauchy scale parameter. Instead, using the 1-stability of Cauchy distribution, Indyk (2000) proves that the median can be employed to estimate the Cauchy scale parameter, and thus the \( \ell_1 \) distances at the projected space \( \mathbb{R}^m \).

Subsequent studies simplified the method proposed by Indyk (2000). Li (2007) shows that \( R \) with Cauchy distribution can be substituted by a sparse \( R \) that has a mixture of symmetric 1-Pareto distribution. A 1-Pareto distribution can be sampled by \( 1/U \), where \( U \) is an independent uniform random variable in \((0, 1)\). This results in a random matrix \( R \) that has the same distribution as described by Equation 6.

The RMI’s two-step procedure is explained using the basic properties of matrix arithmetic and the descriptions given above. Given the projection in Equation 8, the first step of RMI refers to the construction of \( R \): index vectors are the row vectors of \( R \). The second step of the process refers to the construction of \( M' \): context vectors are the row vectors of \( M' \). Using the distributive property of multiplication over addition in matrices,\(^8\)

\(^7\)That is \( E(x) = \infty \), since \( x \) has a Cauchy distribution.

\(^8\)That is, \((A + B)C = AC + BC\).
it can be verified that the explicit construction of M and its multiplication to R can be substituted by a number of summation operations. M can be represented by the sum of unit vectors in which a unit vector corresponds to the co-occurrence of a linguistic entity and a context element. The result of the multiplication of each unit vector and R is the row vector that represents the context element in R—i.e. the index vector. Therefore, M′ can be computed by the accumulation of the row vectors of R that represent encountered context elements, as stated in the second step of the RMI procedure.

2.1 Alternative Distance Estimators
As stated above, Indyk (2000) suggests using the sample median for the estimation of the ℓ1 distances. However, Li (2008) argues that sample median estimator can be biased and inaccurate, specifically if m—i.e. the targeted (reduced) dimensionality—is small. Hence, Li (2008) suggests using the geometric mean estimator instead of the median sample:9

\[ \hat{L}_1(\vec{u}, \vec{v}) = \left( \prod_{i=1}^{m} |u_i - v_i| \right)^{\frac{1}{m}}. \]  (9)

We suggest computing the \( \hat{L}_1(\vec{u}, \vec{v}) \) in Equation 9 using arithmetic mean of logarithm-transformed values of \(|u_i - v_i|\). Therefore, using the logarithmic identities, the multiplications and the power in Equation 9 are, respectively, transformed to a sum and a multiplication:

\[ \hat{L}_1(\vec{u}, \vec{v}) = \exp \left( \frac{1}{m} \sum_{i=1}^{m} \ln(|u_i - v_i|) \right). \]  (10)

Equation 10 for computing \( \hat{L}_1 \) is more plausible for computational implementation than Equation 9 (e.g. the overflow is less likely to happen during the process). Moreover, calculating the median involves sorting an array of real numbers. Thus, computation of the geometric mean in logarithmic scales can be faster than computation of the median sample, especially when the value of m is large.

2.2 RMI’s Parameters
In order to employ the RMI method for the construction of a VSM at reduced dimension and the estimation of the ℓ1 distance between vectors, two model parameters should be decided: (a) the targeted (reduced) dimensionality of the VSM, which is indicated by m in Equation 8 and (b) the number of non-zero elements in index vectors, which is determined by s in Equation 6. In contrast to the classic one-dimension-per-context-element methods of VSM construction,10 the value of m in RPs and thus in RMI is chosen independently of the number of context elements in the model (n in Equation 8).

In RMI, m determines the probability and the maximum expected amount of distortions \( \epsilon \) in the pairwise distance between vectors. Based on the proposed refinements of Indyk (2000, Theorem 3) by Li et al. (2007), it is verified that the pairwise \( \ell_1 \) distance between any \( p \) vectors is approximated within a factor \( 1 \pm \epsilon \), if \( m = O(\log p/\epsilon^2) \), with a constant probability. Therefore, the value of \( \epsilon \) in RMI is subject to the number of vectors \( p \) in the model. For a fixed \( p \), a larger m yields to lower bounds on the distortion with a higher probability. Because a small m is desirable from the computational complexity outlook, the choice of m is often a trade-off between accuracy and efficiency. According to our experiment, m > 400 is suitable for most applications.

The number of non-zero elements in index vectors, however, is decided by the number of context elements n and the sparseness of the VSM \( \beta \) at its original dimension. Li (2007) suggests \( \frac{1}{O(\sqrt{n}\beta)} \) as the value of s in Equation 6. VSMs employed in distributional semantics are highly sparse. The sparsity of a VSM in its original dimension \( \beta \) is often considered to be around 0.0001–0.01. As the original dimension of VSM n is very large—otherwise there would be no need for dimensionality reduction—the index vectors are often very sparse. Similar to m, larger s produces smaller errors; however, it imposes more processes during the construction of a VSM.

2.3 Experimental Evaluation of RMI
We report the performance of the RMI method with respect to its ability to preserve the relative \( \ell_1 \) distance between linguistic entities in a VSM. Therefore, instead of a task-specific evaluation, we show that the relative \( \ell_1 \) distance between a set of words in a high-dimensional word-by-document model remains intact when the model

---

9See also Li et al. (2007, Lemma 5–9).

10That is, n context elements are modelled in an n-dimensional VSM.
is constructed at reduced dimensionality using the RMI technique. We further explore the effect of the RMI’s parameter setting in the observed results.

Depending on the structure of the data that is being analyzed and the objective of the task in hand, the performance of the $\ell_1$ distance for similarity measurement varies from one application to another.\footnote{E.g. see the experiments in Bullinaria and Levy (2007).} The purpose of our reported evaluation, thus, is not to show the superiority of the $\ell_1$ distance (thus RMI) to another similarity measure (e.g. the $\ell_2$ distance or the cosine similarity) and employed techniques for dimensionality reduction (e.g. RI or truncated SVD) in a specific task. If, in a task, the $\ell_1$ distance shows higher performance than the $\ell_2$ distance, then the RMI technique is preferable to the RI technique or truncated SVD. Contrariwise, if the $\ell_2$ norm shows higher performance than the $\ell_1$, then RI or truncated SVD are more desirable than the RMI method.

In our experiment, a word-by-document model is first constructed from the UKWaC corpus at its original high dimension. UKWaC is a freely available corpus of 2,692,692 web documents, nearly 2 billion tokens and 4 million types (Baroni et al., 2009).\footnote{UKWaC can be obtained from http://glo.glo/31fsfE.} Therefore, a word-by-document model constructed from this corpus using the classic one-dimension-per-context-element method has a dimension of 2.69 million. In order to keep the experiments computationally tractable, the reported results are limited to 31 words from this model, which are listed in Table 1.

In the designed experiment, a word from the list is taken as the reference and its $\ell_1$ distance to the remaining 30 words is calculated using the vector representations in the high-dimensional VSM. These 30 words are then sorted in ascending order by the calculated $\ell_1$ distance. The procedure is repeated for all the 31 words in the list, one by one. Therefore, the procedure results in 31 sorted lists, each containing 30 words. Figure 2 shows an example of the obtained sorted list, in which the reference is the word ‘research’.\footnote{Please note that the number of possible arrangements of 30 words without repetition in a list in which the order is important (i.e. all permutations of 30 words) is 30!.}

The procedure described above is replicated to obtain the lists of sorted words from VSMs that are constructed by the RMI method at reduced

| PoS | Words |
|-----|-------|
| Noun | website, students, skills, support, software, research, organisations |
| Adj | online, digital, unique, mobile, sustainable, disabled, innovative |
| Verb | use, visit, improve, provided |

Table 1: Words employed in the experiments.

Figure 2: List of words sorted by their $\ell_1$ distance to the word ‘research’. The distance increases from left to right and top to bottom.

dimensionality, when the method’s parameters—i.e. the dimensionality of VSM and the number of non-zero elements in index vectors—are set differently. We expect the obtained relative $\ell_1$ distances between each reference word and the 30 other words in an RMI-constructed VSM to be the same as the obtained relative distances in the original high-dimensional VSM. Therefore, for each VSM that is constructed by the RMI technique, the resulting sorted lists of words are compared by the sorted lists that are obtained from the original high-dimensional VSM.

We employ the Spearman’s rank correlation coefficient ($\rho$) to compare the sorted lists of words and thus the degree of distance preservation in the RMI-constructed VSMs at reduced dimensionality. The Spearman’s rank correlation measures the strength of association between two ranked variables, i.e. two lists of sorted words in our experiments. Given a list of sorted words obtained from the original high-dimensional VSM ($\text{list}_0$) and its corresponding list obtained from a VSM of reduced dimensionality ($\text{list}_{\text{RMI}}$), the Spearman’s rank correlation for the two lists is calculated by

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)},$$

where $d_i$ is the difference in paired ranks of words in list$_0$ and list$_{\text{RMI}}$, and $n = 30$ is the number of words in each list. We report the average of $\rho$ over the 31 lists of sorted words, denoted by $\bar{\rho}$, to
Figure 3: The $\hat{\rho}$ axis shows the observed average Spearman’s rank correlation between the order of the words in the lists that are sorted by the $\ell_1$ distance obtained from the original high-dimensional VSM and the VSMs that are constructed by RMI at reduced dimensionality using index vectors of various numbers of non-zero elements.

Figure 4 shows the same results as Figure 3, however, in minute detail and only for VSMs of dimension $m \in \{100, 400, 800, 3200\}$. In these plots, squares ($\bullet$) indicate $\hat{\rho}$ while the error bars show the best and the worst observed $\rho$ amongst all the sorted lists of words. The minimum value of $\rho$-axis is set to 0.611, which is the worst observed correlation in the baseline (i.e. randomly generated distances). The dotted line ($\rho = 0.591$) shows the best observed correlation in the baseline and the dashed-dotted line shows the average correlation in the baseline ($\rho = -0.004$). As suggested in Section 2.2, it can be verified that an increase in the dimension of VSMs (i.e. $m$) increases the stability of the obtained results (i.e. the probability of preserving distances increases). Therefore, for large values of $m$ (i.e. $m > 400$), the difference between the best and the worst observed $\rho$ decreases; average correlation $\hat{\rho} \to 1$ and the observed relative distances in RMI-constructed VSMs tend to be identical to those in the original high-dimensional VSM.

Figure 5 represents the obtained results in the same setting as above, however, when the distances are approximated using the geometric mean (Equation 10). The obtained average correlations $\hat{\rho}$ from the geometric mean estimations are almost identical to the median estimations. However, as expected, the geometric mean estimations are more reliable for small values of $m$; particularly, the worst observed correlations when using the geometric mean are higher than those observed when using the median estimator.
The $\ell_1$ distance between context vectors must be estimated using either the median or the geometric mean. The use of the median estimator—for the reasons stated in Section 2.1—is not plausible. On the other hand, the computation of the geometric mean can be laborious as the overflow is highly likely to happen during its computation. Using the value of $\left\lfloor \frac{1}{r} \right\rfloor$ for non-zero elements of index vectors, we know that for any pair of context vectors $\vec{u} = (u_1, \ldots, u_m)$ and $\vec{v} = (v_1, \ldots, v_m)$, if $u_i \neq v_i$ then $|u_i - v_i| \geq 1$. Therefore, for $u_i \neq v_i$, $\ln |u_i - v_i| \geq 0$ and thus $\sum_{i=1}^{m} \ln(|u_i - v_i|) \geq 0$. In this case, the exponent in Equation 10 is a scale factor that can be discarded without a change in the relative distances between vectors. Based on the intuition that the distance between a vector and itself is zero and the explanation given above, inspired by smoothing techniques and without being able to provide mathematical proofs, we suggest estimating the relative distances between vectors using

$$\hat{d}_1(\vec{u}, \vec{v}) = \sum_{i=1}^{m} \ln(|u_i - v_i|). \quad (12)$$

In order to distinguish the above changes in RMI, we name the resulting technique random Manhattan integer indexing (RMII). The experiment described in Section 2.2 is repeated using the RMII method. As shown in Figure 6, the obtained results are almost identical to the observed results when using the RMI technique. While RMI performs slightly better than RMII in lower dimensions, e.g. $m = 400$, RMII shows more stable behaviour than RMI at higher dimensions, e.g. $m = 800$. 

### 4 Comparison of RMI and RI

RMI and RI utilize a similar two-step procedure consisting of the creation of index vectors and the construction of context vectors. Both methods are incremental techniques that construct a VSM at reduced dimensionality directly, without requiring the VSM to be constructed at its original high dimension. Despite these similarities, RMI and RI are motivated by different applications and math-
emathematical theorems. As described above, RMI approximates the $\ell_1$ distance using a non-linear estimator, which has not yet been employed for the construction of VSMs and the calculation of $\ell_1$ distances in distributional approaches to semantics. Moreover, RMI is justified using Cauchy random projections.

In contrast, RI approximates the $\ell_2$ distance using a linear estimator. RI has initially been justified using the mathematical model of the sparse distributed memory (SDM)\(^\text{15}\). Later, Sahlgren (2005) delineates the RI method using the lemma proposed by Johnson and Lindenstrauss (1984)—which elucidates random projections in Euclidean spaces—and the reported refinement in Achlioptas (2001) for the projections employed in the lemma. Although both the RMI and RI methods can be established as $\alpha$-stable random projections—respectively for $\alpha = 1$ and $\alpha = 2$—the methods cannot be compared as they address different goals. If, for a given task, the $\ell_1$ norm outperforms the $\ell_2$ norm, then RMI is preferable to RI. Contrariwise, if the $\ell_2$ norm outperforms the $\ell_1$ norm, then RI is preferable to RMI.

To support the earlier claim that RI-constructed VSMs cannot be used for the $\ell_1$ distance estimation, we evaluate the RI method in the experimental setup that has been used for the evaluation of RMI and RMII. In these experiments, however, we use RI to construct vector spaces at reduced dimensionality and estimate the $\ell_1$ distance using Equation 5 (the standard $\ell_1$ distance definition) and Equation 7 (the median estimator) for $m \in 400, 800$. As shown in Figure 7, the experiments support the theoretical claims.

5 Conclusion

In this paper, we introduce a novel technique, named Random Manhattan Indexing (RMI), for the construction of $\ell_1$ normed VSMs directly at reduced dimensionality. We further suggest the Random Manhattan Integer Indexing (RMII) technique, a computationally enhanced version of the RMI technique. We demonstrated the $\ell_1$ distance preservation ability of the proposed technique in an experimental setup using a word-by-document model. In these experiments, we showed how the variable parameters of the methods, i.e. the number of non-zero elements in index vectors and the

---

\(^{15}\)See Kanerva (1993) for an overview of the SDM model.
dimensionality of the VSM, influence the obtained results. The proposed incremental (and thus efficient and scalable) methods significantly enhance the computation of the $\ell_1$ distances in VSMs.

Acknowledgements

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

References

[Achlioptas2001] Dimitris Achlioptas. 2001. Database-friendly random projections. In Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems, PODS ’01, pages 274–281, New York, NY, USA. ACM.

[Baroni et al.2009] Marco Baroni, Silvia Bernardini, Adriano Ferraresi, and Eros Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. Language Resources and Evaluation, 43(3):209–226.

[Brand2006] Matthew Brand. 2006. Fast low-rank modifications of the thin singular value decomposition. Linear Algebra and its Applications, 415(1):20–30. Special Issue on Large Scale Linear and Nonlinear Eigenvalue Problems.

[Brinkman and Charikar2005] Bo Brinkman and Moses Charikar. 2005. On the impossibility of dimension reduction in $\ell_1$. J. ACM, 52(5):766–788.

[Bullinaria and Levy2007] John A. Bullinaria and Joseph P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior Research Methods, 39:510–526.

[Deerwester et al.1990] Scott C. Deerwester, Thomas K. Landauer, George W. Furnas, and Richard A. Harshman. 1990. Indexing by latent semantic analysis. Journal of the American Society for Information Science, 41(6):391–407.

[Indyk2000] Piotr Indyk. 2000. Stable distributions, pseudorandom generators, embeddings and data stream computation. In Foundations of Computer Science, 2000. Proceedings. 41st Annual Symposium on, pages 189–197.

[Indyk2006] Piotr Indyk. 2006. Stable distributions, pseudorandom generators, embeddings, and data stream computation. J. ACM, 53(3):307–323, May.

[Johnson and Lindenstrauss1984] William Johnson and Joram Lindenstrauss. 1984. Extensions of Lipschitz mappings into a Hilbert space. In Conference in modern analysis and probability (New Haven, Conn., 1982), volume 26 of Contemporary Mathematics, pages 189–206. American Mathematical Society.

[Kanerva et al.2000] Pentti Kanerva, Jan Kristoferson, and Anders Holst. 2000. Random indexing of text samples for latent semantic analysis. In Proceedings of the 22nd Annual Conference of the Cognitive Science Society, pages 103–6. Erlbaum.

[Kanerva1993] Pentti Kanerva. 1993. Sparse distributed memory and related models. In Mohamad H. Hassoun, editor, Associative neural memories: theory and implementation, chapter 3, pages 50–76. Oxford University Press, Inc., New York, NY, USA.

[Ke and Kanade2003] Qifa Ke and Takeo Kanade. 2003. Robust subspace computation using $\ell_1$ norm. Technical Report CMU-CS-03-172, School of Computer Science, Carnegie Mellon University.

[Kwak2008] Nojun Kwak. 2008. Principal component analysis based on 11-norm maximization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 30(9):1672–1680, Sept.

[Lee1999] Lillian Lee. 1999. Measures of distributional similarity. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics, ACL ’99, pages 25–32, Stroudsburg, PA, USA. Association for Computational Linguistics.

[Lenci2008] Alessandro Lenci. 2008. Distributional semantics in linguistic and cognitive research. From context to meaning: Distributional models of the lexicon in linguistics and cognitive science, special issue of the Italian Journal of Linguistics, 201:1–31.

[Li et al.2007] Ping Li, Trevor J. Hastie, and Kenneth W. Church. 2007. Nonlinear estimators and tail bounds for dimension reduction in $L_1$ using cauchy random projections. J. Mach. Learn. Res., 8:2497–2532.

[Li2007] Ping Li. 2007. Very sparse stable random projections for dimension reduction in $L_2$ ($0 < \alpha < 2$) norm. In Proceedings of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '07, pages 440–449, New York, NY, USA. ACM.

[Li2008] Ping Li. 2008. Estimators and tail bounds for dimension reduction in $L_2$ ($0 < \alpha \leq 2$) using stable random projections. In Proceedings of the Nineteenth Annual ACM-SIAM Symposium on Discrete Algorithms, SODA ’08, pages 10–19, Philadelphia, PA, USA. Society for Industrial and Applied Mathematics.

[Martin and Berry2007] Dian I. Martin and Michael W. Berry. 2007. Handbook of latent semantic analysis, chapter Mathematical foundations behind latent semantic analysis, pages 35–55. Ro.
[Sahlgren2005] Magnus Sahlgren. 2005. 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.

[Turney and Pantel2010] Peter D. Turney and Patrick Pantel. 2010. From frequency to meaning: vector space models of semantics. J. Artif. Int. Res., 37(1):141–188, January.

[Weeds et al.2005] Julie Weeds, James Dowdall, Gerold Schneider, Bill Keller, and David Weir. 2005. Using distributional similarity to organise biomedical terminology. Terminology, 11(1):3–4.