An Introduction to Random Indexing

MAGNUS SAHLGREN
SICS, Swedish Institute of Computer Science
Box 1263, SE-164 29 Kista, Sweden
mange@sics.se

Introduction

Word space models enjoy considerable attention in current research on semantic indexing. Most notably, Latent Semantic Analysis/Indexing (LSA/LSI; Deerwester et al., 1990, Landauer & Dumais, 1997) has become a household name in information access research, and deservedly so; LSA has proven its mettle in numerous applications, and has more or less spawned an entire research field since its introduction around 1990. Today, there is a rich flora of word space models available, and there are numerous publications that report exceptional results in many different applications, including information retrieval (Dumais et al., 1988), word sense disambiguation (Schütze, 1993), various semantic knowledge tests (Lund et al., 1995, Karlgren & Sahlgren, 2001), and text categorization (Sahlgren & Karlgren, 2004).

This paper introduces the Random Indexing word space approach, which presents an efficient, scalable and incremental alternative to standard word space methods. The paper is organized as follows: in the next section, we review the basic word space methodology. We then look at some of the problems that are inherent in the basic methodology, and also review some of the solutions that have been proposed in the literature. In the final section, we introduce the Random Indexing word space approach, and briefly review some of the experimental results that have been achieved with Random Indexing.

The word space methodology

The general idea behind word space models is to use distributional statistics to generate high-dimensional vector spaces, in which words are represented by *context vectors* whose relative directions are assumed to indicate semantic similarity. This assumption is motivated by the *distributional hypothesis*, which states that words with similar meanings tend to occur in similar contexts. According to this hypothesis, if we observe two words that constantly occur with the same contexts, we are justified in assuming that they mean similar things. Note that the hypothesis does not require that the words co-

occur with each other; it only requires that the words co-occur with the same *other* words. The viability of the distributional hypothesis has been demonstrated in numerous experiments (Rubenstein & Goodenough, 1965; Charles, 2000).

In the standard word space methodology, the high-dimensional vector space is produced by collecting the data in a co-occurrence matrix F, such that each row F_w represents a unique word w and each column F_c represents a context c, typically a multi-word segment such as a document, or another word. In the former case, where the columns represents documents, we call the matrix a words-by-*documents* matrix, and in the latter case where the columns represents words, we call it a words-by-*words* matrix. LSA is an example of a word space model that uses document-based co-occurrences, and Hyperspace Analogue to Language (HAL; Lund et al., 1995) is an example of a model that uses word-based co-occurrences.[1]

The cells F_{wc} of the co-occurrence matrix record the frequency of co-occurrence of word w and document or word c. As an example, if we use document-based co-occurrences, and observe a given word three times in a given document in the data, we enter 3 in the corresponding cell in the co-occurrence matrix. By the same token, if we use word-based co-occurrences and observe that two given words occur close to each other five times in the data, we enter 5 in the corresponding cell. The frequency counts are usually normalized and weighted in order to reduce the effects of high frequency words, and, in case document-based co-occurrences are used, to compensate for differences in document size.

The point of the co-occurrence matrix is that the rows F_w effectively constitute vectors in a high-dimensional space, such that the elements of the vectors are (normalized) frequency counts, and the dimensionality of the space is determined by the number of columns in the matrix, which is identical to the number of contexts (i.e. words or documents) in the data. We call the vectors *context vectors*, since they represent the contexts in which words have occurred. In effect, the context vectors are representations of the distributional profiles of words, which means that we may define distributional similarity between words in terms of vector similarity. By virtue of the distributional hypothesis, this makes it very straight-forward to compute semantic similarity between words: we simply compare their context vectors using any of a wide range of possible vector similarity measures, such as the cosine of the angles between the vectors, or the City-Block metric.

1. The question whether there is a difference in what kind of information that can be extracted from these two types of co-occurrences has been severly neglected in word space research. We consider this to be highly unfortunate, and believe that this question is of utmost importance.

There are several reasons why this word space methodology has proven to be so attractive – and successful – for a growing number of researchers. The most important reasons are:

- Vector spaces are mathematically well defined and well understood; we know how vector spaces behave, and we have a large set of algebraic tools that we can use to manipulate them.
- The word space methodology makes semantics computable; it allows us to define semantic similarity in mathematical terms, and provides a manageable implementational framework.
- Word space models constitute a purely descriptive approach to semantic modelling; it does not require any previous linguistic or semantic knowledge, and it only detects what is actually there in the data.
- The geometric metaphor of meaning seems intuitively plausible, and is consistent with empirical results from psychological studies.

Problems and solutions

Although theoretically attractive and experimentally successful, word space models are plagued with efficiency and scalability problems. This is especially true when the models are faced with real-world applications and large-scale data sets. The source of these problems is the high dimensionality of the context vectors, which is a direct function of the size of the data. If we use document-based co-occurrences, the dimensionality equals the number of documents in the collection, and if we use word-based co-occurrences, the dimensionality equals the vocabulary, which tends to be even bigger than the number of documents. This means that the co-occurrence matrix will soon become computationally intractable when the vocabulary and the document collection grow.

Another problem with the co-occurrence matrix is that a majority of the cells in the matrix will be zero due to the sparse data problem. That is, only a fraction of the co-occurrence events that are possible in the co-occurrence matrix will actually occur, regardless of the size of the data. A tiny amount of the words in language are distributionally promiscuous; the vast majority of words only occur in a very limited set of contexts. This phenomenon is well known, and is generally referred to as *Zipf's law* (Zipf, 1949). In a typical co-occurrence matrix, more than 99% of the entries are zero.

In order to counter problems with very high dimensionality and data sparseness, most well-known and successful models, like LSA, use statistical dimension reduction techniques. Standard LSA uses truncated Singular Value Decomposition (SVD), which is a matrix factorization technique that can be

used to decompose and approximate a matrix, so that the resulting matrix has much fewer columns – typically only a couple of hundred – and is much denser.[2] It should be noted that SVD is not the only way to achieve this result. There are a number of related dimension reduction techniques that are used in word space research (e.g. principal component analysis and independent component analysis), and they all share the same basic methodology: first sample the data in a standard co-occurrence matrix, and then transform it into a much smaller and denser representation.

Even though they are mathematically well defined and well motivated, there are (at least) three reasons why we would like to avoid using dimension reduction techniques of this type:

- Dimension reduction techniques such as SVD tend to be computationally very costly, with regards to both memory consumption and execution time. For many applications, and especially for large vocabularies and large document collections, it is not practically feasible to compute an SVD.
- Dimension reduction is typically a one-time operation, which means that the entire process of first constructing the co-occurrence matrix and then transforming it has to be done from scratch, every time new data is encountered.[3] The inability to add new data to the model is a serious deficiency, as many applications require the possibility to easily update the model.
- Most importantly, these dimension reduction techniques fail to avoid the initial huge co-occurrence matrix. On the contrary, they *require* initial sampling of the entire data. There are two problems with this. First, it is the initial co-occurrence matrix that is computationally cumbersome. In order to make the models efficient and scalable, this step should be *avoided*, rather than handled by ad hoc solutions. Second, initial sampling of the entire data means that there can be no intermediary results. It is only after we have both constructed *and* transformed the co-occurrence matrix that any processing can begin. If we aim for psychological realism in our word space models, this is a serious liability.

2. See Deerwester et al. (1990) for an introduction to the use of truncated SVD in LSA.
3. It should be noted that there *are* solutions to the problem of adding new data to a reduced space. New data can be "folded" into the already reduced space, but such an operation relies on an old estimate (i.e. the scaling is based on old information, and does not take the new informtaion into account). Thus, there is no guarantee that the old approximation will give reliable estimates for the new information – this problem will be especially severe if the new information is of a topically diverse nature.

Random Indexing

As an alternative to LSA-like models that first construct a huge co-occurrence matrix and then use a separate dimension reduction phase, we have developed an incremental word space model called Random Indexing, based on Pentti Kanerva's work on sparse distributed representations (Kanerva 1988, Kanerva et al., 2000, Kanerva et al., 2001). The basic idea is to *accumulate* context vectors based on the occurrence of words in contexts. This technique can be used with any type of linguistic context, is inherently incremental, and does not require a separate dimension reduction phase.

The Random Indexing technique can be described as a two-step operation:

- First, each context (e.g. each document or each word) in the data is assigned a unique and randomly generated representation called an *index vector*. These index vectors are sparse, high-dimensional, and ternary, which means that their dimensionality (*d*) is on the order of thousands, and that they consist of a small number of randomly distributed +1s and -1s, with the rest of the elements of the vectors set to 0.
- Then, context vectors are produced by scanning through the text, and each time a word occurs in a context (e.g. in a document, or within a sliding context window), that context's *d*-dimensional index vector is added to the context vector for the word in question. Words are thus represented by *d*-dimensional context vectors that are effectively the sum of the words' contexts.

Note that this methodology constitutes a radically different way of conceptualizing how context vectors are constructed. In the "traditional" view, we first construct the co-occurrence matrix and then extract context vectors. In the Random Indexing approach, on the other hand, we view the process backwards, and first accumulate the context vectors. We may then construct a co-occurrence matrix by collecting the context vectors as rows of the matrix.

This means that we can use the Random Indexing procedure to produce a standard co-occurrence matrix F of order $w \times c$ by using *unary* index vectors of the same dimensionality c as the number of contexts, and then collecting the resulting context vectors in a matrix. Such unary index vectors would consist of a single 1 in a different position for each context, and would thus be orthogonal. By contrast, the d-dimensional random index vectors are only *nearly* orthogonal. This means that if we collect the context vectors we produce with Random Indexing in a matrix $F'_{w \times d}$, this matrix will be an *approximation* of the standard co-occurrence matrix $F_{w \times c}$ in the sense that their corresponding rows are similar or dissimilar to the same de-

gree, but with $d << c$. In this way, we can achieve the same effect as is done in LSA by the use of SVD: transforming the original co-occurrence counts into a much smaller and denser representation.

Random Indexing can be motivated through an observation made by Hecht-Nielsen (1994), who demonstrated that there are many more nearly orthogonal than truly orthogonal directions in a high-dimensional space. This means that we can approximate orthogonality by simply choosing random directions in the high-dimensional space. This near-orthogonality of the random vectors is the key to a family of dimension reduction techniques that includes methods such as Random Projection (Papadimitriou et al., 1998), Random Mapping (Kaski, 1999), and Random Indexing. These methods rest on the same insight – the Johnson-Lindenstrauss lemma (Johnson & Lindenstrauss, 1984) – that states that if we project points in a vector space into a randomly selected subspace of sufficiently high dimensionality, the distances between the points are approximately preserved. Thus, the dimensionality of a given matrix F can be reduced by multiplying it with (or projecting it through) a random matrix R:

$$F_{w \times d} R_{d \times k} = F'_{w \times k}$$

Obviously, the choice of random matrix R is an important design decision for dimension reduction techniques that rely on the Johnson-Lindenstrauss lemma. If the random vectors in matrix R are orthogonal, so that $R^{T}R = I$, then $F = F'$; if the random vectors are nearly orthogonal, then $F \approx F'$ in terms of the similarity of their rows. A very common choice for matrix R is to use Gaussian distribution for the elements of the random vectors. However, Achlioptas (2001) has shown that much simpler distributions – practically all zero mean distributions with unit variance – give a mapping that satisfies the lemma. Random Indexing uses a variant of Achlioptas' proposal for the distribution of non-zero elements in the random index vectors.

Compared to other word space methodologies, the Random Indexing approach is unique in the following four ways:

- First, it is an incremental method, which means that the context vectors can be used for similarity computations even after just a few examples have been encountered. By contrast, most other word space models require the entire data to be sampled before similarity computations can be performed.
- Second, the dimensionality d of the vectors is a *parameter* in Random Indexing. This means that d does not change once it has been set; new data increases the values of the elements of the context vec-

tors, but never their dimensionality. Increasing dimensionality can lead to significant scalability problems in other word space methods.

- Third, Random Indexing uses "implicit" dimension reduction, since the fixed dimensionality d is much lower than the number of contexts c in the data. This leads to a significant gain in processing time and memory consumption as compared to word space methods that employ computationally expensive dimension reduction algorithms.
- Fourth, Random Indexing can be used with any type of context. Other word space models typically use either documents or words as contexts. Random Indexing is not limited to these naive choices, but can be used with basically any type of contexts.

Results

Random Indexing has been empirically validated in a number of experiments. Kanerva et al. (2000) used Random Indexing with document-based co-occurrence statistics to solve the synonym-finding part of the TOEFL, in which the subject is asked to choose a synonym to a given word out of four provided alternatives. Karlgren and Sahlgren (2001) used word-based co-occurrence statistics to enhance the performance of Random Indexing in the same task to 64.5% – 67% correct answers, which is comparable to results reported for LSA (64.4%), and results for foreign applicants to U.S. colleges (64.5%). The performance of Random Indexing was even further enhanced by using lemmatization of the text data (and thereby reducing the number of unique word types), which produced a top score of 72% correct answers.

Sahlgren & Cöster (2004) used Random Indexing to improve the performance of text categorization. The idea was to produce concept-based text representations based on Random Indexing, and to use those representations as input to a support vector machine classifier. The results using the concept-based representations were comparable to standard text representations when counting all 90 categories (around 82%), but were slightly better when only the ten largest categories were considered (around 88%).

Sahlgren & Karlgren (2005) demonstrated that Random Indexing can be applied to multilingual data. In their experiments, they demonstrated how bilingual lexica can be extracted using Random Indexing applied to parallel data. The overlap between the extracted lexica and manually compiled gold standard lexica was around 60%.

More references to experiments using Random Indexing can be found at: `http://www.sics.se/~mange/publications.html`

References

Achlioptas, D.; Database-friendly random projections. *Symposium on principles of database system*s. 2001.

Charles, W.; Contextual correlates of meaning. *Applied psycholinguistics*. 21, Cambridge University Press, 2000.

Deerwester, S.; S. Dumais; G. Furnas; T. Landauer; R. Harshman; Indexing by Latent Semantic Analysis. *Journal of the society for information science*. 41(6), 1990.

Dumais, S.; G. Furnas; T. Landauer; S. Deerwester; Using Latent Semantic Analysis to improve access to textual information. *Proceedings of CHI'88*. New York: ACM, 1988.

Hecht-Nielsen, R.; Context vectors; general purpose approximate meaning representations self-organized from raw data. In Zurada, J. M.; R. J. Marks II; C. J. Robinson; *Computational intelligence: imitating life*. IEEE Press, 1994.

Johnson, W. B. & J. Lindenstrauss; Extensions to Lipshitz mapping into Hilbert space. *Contemporary mathematics*. 26, 1984.

Kanerva, P.; *Sparse distributed memory*. The MIT Press, 1988.

Kanerva, P.; J. Kristofersson; A. Holst; Random Indexing of text samples for Latent Semantic Analysis. *Proceedings of the 22nd annual conference of the cognitive science society*. New Jersey: Erlbaum, 2000.

Kanerva, P.; G. Sjödin; J. Kristofersson; R. Karlsson; B. Levin; A. Holst; J. Karlgren; M. Sahlgren; Computing with large random patterns. In Uesaka, Y.; P. Kanerva; H. Asoh; *Foundations of real-world intelligence*. Stanford: CSLI Publications, 2001.

Karlgren, J. & M. Sahlgren; From words to understanding. In Uesaka, Y.; P. Kanerva; H. Asoh; *Foundations of real-world intelligence*. Stanford: CSLI Publications, 2001.

Kaski, S.; Dimensionality reduction by random mapping: fast similarity computation for clustering. *Proceedings of the IJCNN'98*. IEEE Service Center, 1998.

Landauer, T. & S. Dumais; A solution to Plato's problem: the Latent Semantic Analysis theory for acquisition, induction and representation of knowledge. *Psychological review*. 104(2), 1997.

Lund, K; C. Burgess; R. A. Atchley; Semantic and associative priming in high-dimensional semantic space. *Proceedings of the 17th annual conference of the cognitive science society*. New Jersey: Erlbaum, 1995.

Papadimitriou, C. H.; P. Raghavan; H. Tamaki; S. Vempala; Latent Semantic Indexing: a probabilistic analysis. *Proceedings of the 17th ACM symposium on the principles of database systems*. ACM Press, 1998.

Rubenstein, H. & J. Goodenough; Contextual correlates of synonymy. *Commun. ACM*, 8(10), 1965.

Sahlgren, M. & R. Cöster; Using bag-of-concepts to improve the performance of support vector machines. *Proceedings of COLING 2004*. Geneva, 2004.

Sahlgren. M. & J. Karlgren; Automatic bilingual lexicon acquisition using Random Indexing of parallel corpora. *Journal of Natural Language Engineering, Special Issue on Parallel Texts*. June, 2005.

Schütze, H.; Word space. In S. Hanson; J. Cowan; C. Giles; *Advances in Neural Information Processing Systems 5*. Morgan Kaufmann Publishers, 1993.

Zipf, G. K.; *Human behavior and the principle of least effort*. Addison-Wesley, 1949.