Random Indexing of Text Samples for Latent Semantic Analysis

Pentti Kanerva Jan Kristoferson Anders Holst
kanerva@sics.se, janke@sics.se, aho@sics.se
RWCP Theoretical Foundation SICS Laboratory
Swedish Institute of Computer Science, Box 1263, SE-16429 Kista, Sweden

Latent Semantic Analysis is a method of computing high-dimensional semantic vectors, or context vectors, for words from their co-occurrence statistics. An experiment by Landauer & Dumais (1997) covers a vocabulary of 60,000 words (unique letter strings delimited by word-space characters) in 30,000 contexts (text samples or "documents" of about 150 words each). The data are first collected into a $60{,}000 \times 30{,}000$ words-by-contexts co-occurrence matrix, with each row representing a word and each column representing a text sample so that each entry gives the frequency of a given word in a given text sample. The frequencies are normalized, and the normalized matrix is transformed with Singular-Value Decomposition (SVD) reducing its original 30,000 document dimensions into a much smaller number of latent dimensions, 300 proving to be optimal. Thus words are represented by 300-dimensional semantic vectors.

The point in all of this is that the vectors capture meaning. Landauer and Dumais demonstrate it with a synonym test called TOEFL (for "Test Of English as a Foreign Language"). For each test word, four alternatives are given, and the "contestant" is asked to find the one that's the most synonymous. Choosing at random would yield 25% correct. However, when the semantic vector for the test word is compared to the semantic vectors for the four alternatives, it correlates most highly with the correct alternative in 64% of the cases. However, when the same test is based on the 30,000-dimensional vectors before SVD, the result is not nearly as good: only 36% correct. The authors conclude that the reorganization of information by SVD somehow corresponds to human psychology.

We have studied high-dimensional random distributed representations, as models of brainlike representation of information (Kanerva, 1994; Kanerva & Sjödin, 1999). In this poster we report on the use of such a representation to reduce the dimensionality of the original words-by-contexts matrix. The method can be explained by looking at the $60{,}000 \times 30{,}000$ matrix of frequencies above. Assume that each text sample is represented by a 30,000-bit vector with a single 1 marking the place of the sample in a list of all samples, and call it the sample's *index vector* (i.e., the nth bit of the index vector for the nth text sample is 1—the representation is unitary or local). Then the words-by-contexts matrix of frequencies can be gotten by the following procedure: every time that the word w occurs in the nth text sample, the nth index vector is added to the row for the word w.

We use the same procedure for accumulating a words-by-contexts matrix, except that the index vectors are not unitary. A text-sample's index vector is "small" by comparison—we have used 1,800-dimensional index vectors—and it has several randomly placed -1s and 1s, with the rest 0s (e.g., four each of -1 and 1, or eight non-0s in 1,800, instead of one non-0 in 30,000 as above). Thus, we would accumulate the same data into a $60{,}000 \times 1{,}800$ words-by-contexts matrix instead of $60{,}000 \times 30{,}000$.

Our method has been verified with different data, a ten-million-word "TASA" corpus consisting of a 79,000-word vocabulary (when words are truncated after the 8th character) in 37,600 text samples. The data were accumulated into a $79{,}000 \times 1{,}800$ words-by-contexts matrix, which was normalized by thresholding into a matrix of -1s, 0s, and 1s. The unnormalized 1,800-dimensional context vectors gave 35–44% correct in the TOEFL test and the normalized ones gave 48–51% correct, which correspond to Landauer & Dumais' 36% for their normalized 30,000-dimensional vectors before SVD, for a different corpus (see above). Our words-by-contexts matrix can be transformed further, for example with SVD as in LSA, except that the matrix is much smaller.

Mathematically, the 30,000- or 37,600-dimensional index vectors are orthogonal, whereas the 1,800-dimensional ones are only nearly orthogonal. They seem to work just as well, in addition to which they are more "brainlike" and less affected by the number of text samples (1,800-dimensional index vectors can cover a wide-ranging number of text samples). We have used such vectors also to index words in narrow context windows, getting 62–70% correct, and conclude that random indexing deserves to be studied and understood more fully.

Acknowledgments. This research is supported by Japan's Ministry of International Trade and Industry (MITI) under the Real World Computing Partnership (RWCP) program. The TASA corpus and 80 TOEFL test items were made available to us by courtesy of Professor Thomas Landauer, University of Colorado.

References

Kanerva, P. (1994). The Spatter Code for encoding concepts at many levels. In M. Marinaro and P. G. Morasso (eds.), *ICANN '94, Proc. Int'l Conference on Artificial Neural Networks* (Sorrento, Italy), vol. 1, pp. 226–229. London: Springer-Verlag.

Kanerva, P., and Sjödin, G. (1999). Stochastic Pattern Computing. *Proc. 2000 Real World Computing Symbosium* (Report TR-99-002, pp. 271–276). Tsukuba-city, Japan: Real World Computing Partnership.

Landauer, T. K., and Dumais, S. T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review* 104 (2):211–240.