Random Manhattan Indexing

A Randomized Scalable Method for Semantic Similarity Measurement in L_1 Normed Spaces

Behrang Q. Zadeh
Insight Centre for Data Analytics
National University of Ireland, Galway
behrang.qasemizadeh@insight-centre.org

Abstract

Vector space models are well-defined mathematical representation framework that have been widely used in text analytics. In order to deliver a solution for problems that require a minimal level of text understanding, in these models, text units are represented by high-dimensional vectors. The constructed vector spaces are endowed with a norm structure and a distance formula is employed to compute the similarity of vectors, thus, the similarity of the text units that they represent. The high dimensionality of the vectors, however, is a barrier to the performance of these models. We introduce Random Manhattan Indexing (RMI) for the construction of L_1 normed vector space models of semantics at reduced dimension. RMI is a two-step incremental method of vector space construction that employs a sparse stable random projection to achieve its objective.

1. Motivation

Distributional approaches to semantics tie the meaning of text units to their usage context. These methods attempt to quantify the meaning of text units by investigating their distributional similarities. A vector space is an algebraic structure that can be employed to represent such distributional similarities. In this model, the relative proximity of vectors to one another interprets the meaning of text units that they represent. However, as the number of text units that are being modelled in a VSM increases, the number of contexts that are required to be utilized to capture their meaning escalates. This phenomenon is explained using power-law distributions of text units in contexts. For example, Zipf's law states that most words are rare, while few words are used frequently. As a result, extremely high-dimensional vectors, which are also sparse, represent text units. The high-dimensionality of the vectors results in obstacles, which are known as *the curse of dimensionality*. A dimension reduction method is often required to alleviate these problems.

2. Random Manhattan Indexing

We employ stable random projections and introduce the RMI technique for the incremental construction of vector spaces in L_1 normed spaces. In this method, the dimension of the vector space is fixed, independent of the text-data, and known prior to the task of vector space construction. The method, thus, is an excellent choice for processing big text-data, at large scale such as web.

RMI employs a two-step procedure. First, context elements are assigned to *index vectors*. Index vectors are unique and generated randomly such that entries r_i of index vectors have the following distribution:

$$r_i = \begin{cases} \frac{-1}{U_1} & \textit{with probability } \frac{s}{2} \\ 0 & \textit{with probability } 1-s \\ \frac{1}{U_2} & \textit{with probability } \frac{s}{2} \end{cases}, \qquad (1)$$

where U_1 and U_2 are independent uniform random variables in (0,1). In the second step, each text unit is assigned to a context vector $\vec{v}_c$ where initially all the elements of $\vec{v}_c$ are set to 0. For each encountered co-occurrence of a text unit and a context element, the context vector of the text unit is accumulated by the index vector $\vec{r}_i$ of the context element, i.e. $\vec{v}_c = \vec{v}_c + \vec{r}_i$. The result is a VSM at reduced dimension that can be used to estimate the pairwise L_1 distance between text units in the model. In the constructed vector space, the logarithmic geometric mean can be used to estimate the L_1 distance between vectors:

$$\hat{L}_1(\vec{u}, \vec{v}) = \exp(\frac{1}{m}\sum_{i=1}^{m} \ln(|u_i - v_i|)). \qquad (2)$$

The proposed method can be verified mathematically and has been validated by a set of experiments [2]. In addition, a computationally enhanced variation of the RMI method can be found in [1].

Acknowledgment

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

References

[1] B. Q. Zadeh and S. Handschuh. Random manhattan indexing. In *Proceedings of 25th International Workshop on Database and Expert Systems Applications(DEXA)*, 2014.

[2] B. Q. Zadeh and S. Handschuh. Random manhattan integer indexing: Incremental l1 normed vector space construction. In *Empirical Methods on Natural Language Processing*, 2014.