THE UNIVERSITY OF MEMPHIS | Office of Technology Transfer

Systematically Evaluate Semantic Differences between Semantic Spaces

The University of Memphis seeks a commercial partner to bring an exciting new technology to market. Specifically, a method that systematically evaluates semantic differences between semantic spaces, developed by the Institute for Intelligent Systems at the University of Memphis (https://sites.google.com/a/iis.memphis.edu/main/), may be of very high value to Google's search engine and other text analysis and optimization tasks. Further, sample implementation with the Google framework is available for demonstration at http://about.dsspp.com.

A pending patent application is available for licensing. Contact Hai Trieu at 901-678-1712, hhtrieu@memphis.edu

What makes this unique?

This process provides a method for evaluating semantic spaces within a unified theoretical framework. Researchers and companies can objectively find the best semantic spaces engines to evaluate their content. This allows for customized creation/selection of a semantic space that is the most effective for the current project's needs.

Potential Application

This process can be applied to optimize (or individualize) search functions. Terms entered into a search engine can be automatically expanded to similar neighbor terms. This focuses the search into relevant areas. Alternatively this could improve natural language processing abilities.
Examples of relevant applications

- Focusing search results.
- Pinpoint relevant items to better customize delivery of web advertising.

Importance of the Invention

Because semantic spaces are a core component of advanced information retrieval (AIR), advanced learning environments (ALE), and computerized text analysis (CTA), this invention will improve the efficiency and quality of these systems. The invention is a major innovation in advancing research in semantic spaces. It will significantly broaden the current capabilities of researchers in evaluating written texts.

Potential of the Invention

Given the theoretical framework and basic methods provided in the previous sections, this invention will be able to solve the challenges in semantic spaces. In the section below, we outline three types of applications: (1) computer implementation; (2) understanding the properties of the similarity curves; and (3) evaluation of semantic spaces.

Invention Description

This invention is based on work done by the inventor (Hu et al., 2005), who outlined a mathematical framework that characterizes most semantic spaces such as LSA, HAL, NLS, and Topics Model. The basic idea is to formulate semantic spaces as an algebraic structure with similarity measures. The most important concept in this theory is the concept of induced semantic structure (ISS) for a semantic space. With ISS, comparisons can be made between semantic spaces. Such comparisons are operationally defined within the framework, and they are numerical. Based on these measures, we achieve our goals in this invention.

Next we provide some basic assumptions and definitions for the theoretical framework of semantic spaces. We first briefly outline some basic concepts from our previous work and then turn to the description of the invention.

Theoretical Foundation

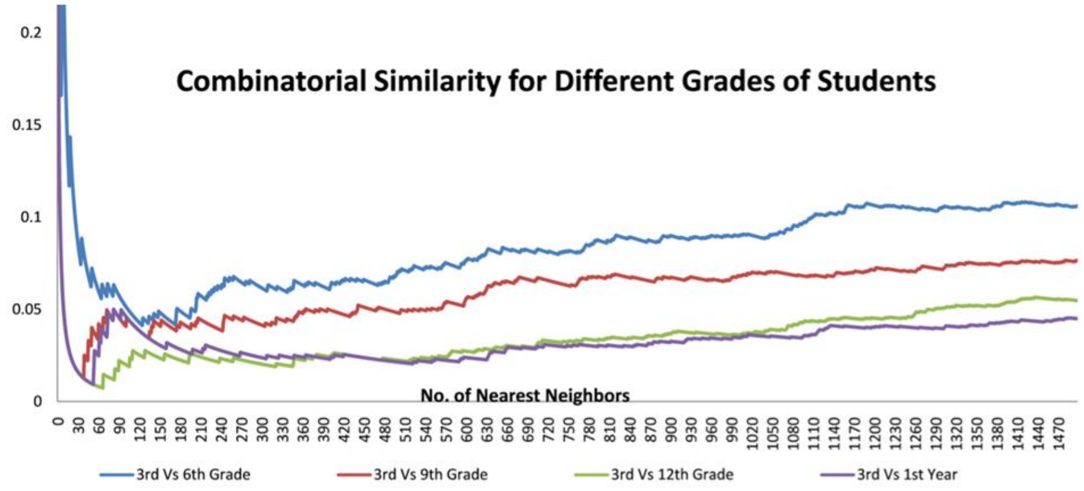
Hu et al.'s (2005) framework of semantic spaces proposes: (1) a general definition of semantic space, characterizing the most widely used semantic spaces such as LSA, HAL, Topics Model, and NLS; (2) the induced semantic structure of a semantic space; and (3) three levels of measurements between semantic spaces.

Applications of the Theoretical Framework

In the previous section, we introduced one major concept of induced semantic structure and three levels of similarity measures: combinatorial, permutational, and quantitative similarities between items and between spaces (statistical properties of sets). This invention is based on the formal framework of induced semantic structure and these measures. To illustrate this framework, we give a simple example of measuring semantic similarity at the item level and outline detailed procedures for measuring semantic similarity at the space level.

Assume that we have two LSA spaces L_1 and L_2 with a common set of words W. Two matrices can be obtained by considering near neighbors for all words in W: S_1 and S_2, such that each corresponds to the similarity measure among all words in W for the two LSA spaces L_1 and L_2. Thus S_1 and S_2 are square and symmetric, and have dimension [W]. Note that these two similarity matrices contain all necessary information needed for combinatorial, permutational, and quantitative similarities measures.

In this example we compute combinatorial, permutational, and quantitative similarities for the word "life." The chart below lists near neighbors for several LSA spaces. We computed similarities with the value T = 1…1500 and observed that the meaning of "life" is most similar between third grade and sixth grade in the TASA corpus.



It is very important to note that:

- Values computed from the equations for combinatorial, permutational, and quantitative similarity are functions of the value *T*. By varying *T* a more optimal structure may be induced for a particular semantic space.

- Use of combinatorial, permutational, and quantitative similarity is at the smallest resolution of the LSA spaces, namely, the items. They can also be applied to the levels of phrases, sentences, or documents (e.g. , one may evaluate the meaning of the document across different spaces).

Measuring semantic differences for a collection of items between semantic spaces
In the previous section, the descriptions for combinatorial, permutational, and quantitative similarity only provide similarity measures between items. In the example, we used the descriptions to measure the meaning of "life" between different semantic spaces. If a collection of items is considered, numerical values can be obtained for each of the terms. For example, if *W* is a collection of furniture items, then the combinatorial measure would be a collection of combinatorial similarities for those items between two target semantic spaces. Statistical properties can be obtained to measure the differences. Although *W* is only a collection of items, it is very important to know the following:

- If *W* is the overlap of all items in the target semantic spaces, statistical properties of combinatorial, permutational, and quantitative similarity will reflect semantic differences between the target semantic spaces.

- *W* can be a collection of phrases, sentences, and documents. The invention makes it possible to examine the difference between two completely different semantic spaces with the same corpus at the level of documents.

Computer Implementation
The most essential component of this invention is the use of high capacity computers to carry out a huge amount of computation. We use Latent Semantic Analysis (LSA) as an example below; the same procedure can be applied to any other vector-based semantic encoding methods.

By using the computing power available (High Performance Computing facility), we will divide "computer implementation" into a series of small projects:

1. ***Improve LSA***: Create an LSA utility for the super-computers. LSA is the most studied type of semantic space that has been used in advanced learning environments, particularly in information retrieval, intelligent tutoring systems, and text analysis. This task will be a combination of migrating existing LSA algorithms and creating new components that are parameter controlled. With the new LSA generating software, LSA spaces can be generated in a matter of minutes based on any specification of the seven steps (cf. Section 2.3).

2. ***Generate nearest neighbors for semantic spaces***: Create programs that can generate neighbors of any given item x. This program will require heavy computing. It is actually creating an *N-by-N* matrix whose entries are pair-wise similarity measures within a given layer of semantic space, where *N* is the number of items (of the given layer) in the given semantic space. Such a matrix will be used to generate the induced semantic structure and the three levels of similarity measures.

3. ***Compute similarity measures***: Creative programs that can compute the three levels of similarity measures for any two elements in the same layer of either the same semantic space or different semantic spaces.

4. ***Implement other semantic spaces***: Create software to generate semantic spaces such as HAL (Burgess, Livesay , & Lund, 1996). NLS (Cai et al ., 2004), and Topics Model (Griffiths et al., 2007; Steyvers & Griffiths, 2007).

The above four projects are computational in nature and require a large number of computations.

The Inventor

Xiangen Hu, Ph.D. is a professor in the Department of Psychology at The University of Memphis. Dr. Hu received his Master (applied mathematics) from Huazhong University of Science and Technology in 1985, Master (Social Sciences) in 1991 and Ph.D. (Cognitive Sciences) in 1993 from the University of California, Irvine. He joined the University of Memphis in 1993.

Dr. Hu’s primary research areas include mathematical Psychology, Research Design and Statistics, and Cognitive Psychology. More specific research interests include General Processing Tree (GPT) models, categorical data analysis, knowledge representation, computerized tutoring, and advanced distributed learning.