Shadow Searching in Concept Cloud

Mishra Pavan <pavanmishra@gmail.com>, Surve Akshay <akshay.surve@gmail.com>

Introduction

We propose a Concept Cloud for recursive representation of concepts representing the art world. The collection of concepts is then used for querying the existing understanding of art and discovering new ones.

Vocabulary

Concept

Concepts are repetitive abstract patterns of cognition. They may refer to other concepts and may also embody others. Thus concept is a recursive construct wherein it can be defined in terms of other concepts. Concepts are also related to each other evolving from the recursive definition of concept.

Concept Cloud

Concept Cloud is a hyperspace comprised of related concepts representing the domain separated by semantic distance.

Advantages of representing the domain recursively in a Concept Cloud:

- Abstract representation: Since the actual total structure of the data is not known until run-time and grows, the data must be represented by an abstraction, such as an abstract class or interface.
- Base case(s): These represent the "end" of the pattern. They are the termination point(s) of the data structure.
- Inductive case(s): These represent the on-going, "interior" portion of the repetitive pattern. They embody the ability to represent the data structure as a simple connection between abstractly equivalent entities.

Relationship

The relative association between the concepts is determined by the semantic concept distance which is discussed below. These associations can further be enriched with annotations to represent definite assertions between the concepts. The annotations also contribute to the strength of the relation and as such are among the metrics used to calculate the concept distance.

Semantic Concept Distance

The semantic concept distance is a quantifiable measure of the association between the concepts. It depends on occurrence mapping of other concepts in recursion, on annotated relationships if defined and on any other metrics from the domain.
This measure is of prime importance for functioning and optimization of Concept Inference Engine. The distance keeps changing as new concepts and relationships are discovered. We use the Tversky's semantic distance formula.

Concept Inference Engine

The inference engine is the intelligent component in the discussion, as such it is involved in the building of the concept cloud, mapping the model concept(searching) and discovering of new concepts. The inference engine makes the use of concept and semantic distance between them for all its needs. We propose the Concept Inference Engine to be based on fuzzy clustering.

Pseudocode

Semantic Concept Distance

$$S(a, b) = \frac{|A \cap B|}{|A \cap B| + \alpha(a, b)\,|A - B| + (1 - \alpha(a, b))|B - A|}$$

S(a, b) is the similarity between two arbitrary objects, a, b
A and B are feature sets of a, b respectively
α is a real no. $\ni 0 \leq \alpha \leq 1$

Building and Searching in the Concept Cloud

The approach in building the Concept Cloud would depend on the information source under use.
If the source is generic then it would require training as to what belongs to the domain and what not. This would be supplemented by the supervised learning system. However if the source is domain specific, it can analyze the source and develop the domain concepts. Concepts discovered earlier are then automatically clustered using semantic concept distance.

While searching the very nature of the recursive representation of concepts in the cloud yields for inexact mapping. The model concept being mapped is projected onto the concept cloud. Only those concepts are considered which fall within the pre-specified semantic concept distance. The concepts are ranked with respect to the concept distance from the model concept. So the concept nearer to the model concept is ranked higher and vice-versa.

Clustering in the Concept Cloud (Concept Inference Engine) can be based on fuzzy clustering:

$$\forall x \left(\sum_{k=1}^{\text{num. clusters}} u_k(x) = 1 \right).$$

The centroid of a cluster is the mean of all points, weighted by their degree of belonging to the cluster

$$\text{center}_k = \frac{\sum_x u_k(x)^m x}{\sum_x u_k(x)^m}.$$

The degree of belonging is related to the inverse of the distance to the cluster center:

$$u_k(x) = \frac{1}{d(\text{center}_k, x)},$$

Discovering new concepts

As new concepts and relationships are mapped, the clusters in the cloud are rearranged. From time to time new clusters will be formed and these clusters will develop into whole new concepts themselves. The system can then be assisted to accept these new discoveries or rely on certain threshold limit as to incorporate new concepts into the system.

$$C = \{C_i \, \forall C_i : S(C, C_i) \leq \alpha\}$$

α - is the threshold limit