# Abstract

The ability to identify analogies and correspondences is one of the fascinating aspects of intelligence. It allows learning across different situations, systems and domains, where the common base to learning is not trivial or immediate. The research in cognitive science has acknowledged the significance of analogy making to human thinking. Several previous works on analogy making suggested computational mechanisms for constructing detailed mapping that connects corresponding ingredients across a given pair of analogized systems.

In this work, we introduce a new approach to understanding, identifying and forming analogies and correspondences. In distinction from previous works on analogies, our approach and the computational methods derived from it are applicable to real world problems, such as the task of identification of corresponding topics in texts on different domains. This work thus bridges between cognitive observations regarding analogy making, which inspired this work but provided no concrete utilizable computational recipes, with techniques that have been proven efficient in processing real world data.

The methods introduced in this work extend the well known data clustering problem. The key mechanism used to identify correspondences through clustering is directing corresponding data elements, from different subsets of elements each representing one of the systems between which the correspondence is being drawn, to be included in the same cluster. The straightforward application of a standard clustering technique would not address well this target: a standard clustering method would often produce clusters with elements of only one of the representative subsets, particularly when these subsets are relatively homogenous. Our methods, however, are specifically designed to cluster together corresponding elements from both subsets while neutralizing the impact of homogeneity within each subset.

The first method that we introduce, termed *coupled clustering*, addresses the problem of partitioning two representative element subsets. This method extends an axiomatic framework of similarity-based data clustering (Puzicha, Hofmann & Buhmann, 2000). The other method, *cross-partition clustering*, modifies and generalizes the coupled clustering setting along several aspects: it is based on vectorial representation of the given data rather than on pairwise proximity values, it produces soft (probabilistic) rather than deterministic partitioning and it allows revealing correspondences across more than two subsets. The cross-partition clustering method is based on the *information bottleneck*

(Tishby, Pereira & Bialek, 1999) and *information distortion* (Gedeon, Parker & Dimitrov, 2003) methods, which are grounded on information theoretic approach to data clustering.

The setting underlying our approach is considerably different than previous views of analogy making. The two methods that we introduce ascribe the correspondence being formed to a counterbalance between different factors. In coupled clustering, these factors are shared pairwise similarity (across subsets) and prominence of the formed cluster in each one of both analogized subsets. In the cross partition method, the underlying factors are communal feature distribution patterns versus the independence of these patterns on the pre-partition of the data to distinct subsets between which a correspondence is revealed.

Both methods were developed using general formulation. They are capable of identifiable correspondences drawn across any sets of data elements that are represented through feature vectors or a similarity matrix, regardless of the source of the data. Hence in principle, they are applicable to a large variety of problems and domains. In this work both methods were applied successfully to synthetic and textual data.

The textual experiments addressed the task of identifying corresponding sub-topics across related but distinct domains, each represented through a set of domain-related keywords extracted from appropriate corpora. The similarity measure and vectorial representations required as input to our framework were compiled based on word co-occurrence statistics. Most experiments were focused on identifying correspondence between different religions: Buddhism, Christianity, Hinduism, Islam and Judaism. With no prior specialization or training in the study of religions, our methods identified analogous factors shared by several religions in varying levels of resolution: “spiritual” versus “practical” dimensions in a coarse view and aspects such as “sacred writings”, “rite and festivals” and “sin and suffering” in more detailed level. These findings are in apparent agreement with comparative religion studies that were based on a comparable approach. For the purpose of systematic evaluation, we have measured the overlap between our outcome and religion-related term clusters provided by experts. The match between the experts’ clusters and the outcome of our method was very close to the level of agreement between the experts.