Chapter 6:

Discussion and Further Directions

In this work, we have defined and studied a new unsupervised computational learning task: automated identification of correspondences across a dataset pre-divided into several – two or more – subsets. We have developed methods that accomplish this task and we have demonstrated them on synthetic data as well as on real world data extracted from un-annotated texts. As our methods are introduced using general formulation that does not depend on their application to texts, our approach is potentially utilizable for a wide variety of real-world problems. At the same time, it opens new perspectives on the analogy making task, which has been typically associated with cognitive concepts and mental processes such as discovery and creation.

Our work extends the data clustering task, in a way that enables coping with analogy or correspondence identification. Correspondences are identified by way of assigning together corresponding elements from different subsets into the same cluster. As we have emphasized (Section 3.1), the straightforward application of a standard clustering technique would not address well the above task. This is particularly true when each of the pre-given subsets given as input is relatively homogenous and overall not very similar to the other subsets. In such cases, a standard clustering method would tend to produce clusters with elements restricted to a single subset. Our results, however, demonstrate the capability of modified clustering techniques to reveal correspondences between the subsets as required, rather than subset-specific themes.

As mentioned (Subsection 2.1.1.2), the data clustering problem is formally ill posed. In practice, the quality of a proposed solution for a specific is assessed in terms of the requirements of the specific application. Our modified data-clustering problem is subject to the same type of ambiguity and, in fact, the potential source of ambiguity is even heaped on. While ambiguity in the original data clustering task results from the lack of a definite criterion for how elements should be grouped together, the extended task adds on top of that a potential ambiguity regarding constructing the matches across two or more given subsets. In spite of being formally ill posed, the data clustering task has been studied extensively. We hope our new task would be recognized as a useful tool that deserves further study just as the standard data-clustering notion on which it is based.

We have developed two different data-clustering based computational methods that identify correspondence across several given subsets of data elements. The first of which, *coupled clustering* (Chapter 4), is based on a recent cost-based pairwise clustering framework. In this setting, the distributional data representation that is typically given needs conversion to pairwise similarities. The second method, *cross partition clustering* (Chapter 5), extends clustering methods grounded on information theoretic accounts and is directly based on co-occurrence distribution. Each of these two methods bases its strategy on few principles, pertaining to the essence of the task of identifying correspondences.

There are several advantages to starting from studying a setting based on deterministic ("hard") data-clustering, as we have done in Chapter 4, rather than the probabilistic or "soft" variations. Deterministic clustering is technically and conceptually simpler and it constructs more definite and easily interpretable clustering configurations. The coupled clustering method restricts the similarity values generally considered by pairwise clustering methods to similarities of elements that are not in the same subset ("between-subset similarities"). It is thus guided by the working assumption that information within a subset is not supposed to impact directly the correspondences formed across subsets but, rather, the information resulting from comparing two subsets, i.e., similarities between members of the different subsets, should be the factor to consider in constructing such correspondences. This assumption motivates the main original mechanism underlying the coupled clustering method: the cost-function that we have proposed (H^3, Eq. 4.14), which incorporates two complementary principles. The first one is the underlying principle of pairwise clustering in general:

> *A cluster should contain elements that are similar to one another.*

The second principle turned to be the underlying idea of the coupled clustering method:

> *In order to be formed, a cluster must exceed some level of prominence in both subsets (as opposed to a cluster that is overall more prominent, but most of its members are concentrated in one of the subsets).*

This second direction is realized through a geometric-mean term that is used for calculating average similarity in each cluster.

In summary, the coupled clustering method is a rather straightforward elaboration on the standard cost-based pairwise clustering setting, which only restricts the collection of similarity values under consideration to the collection of between-subset similarities. The essential drawback of the coupled clustering method might lie in its presumed working assumption. It is probable that similarities between elements of distinct subsets are more important for the emerged correspondence, but the policy of restricting the attention to these similarities is, in retrospective, just a preliminary rough direction. The coupled clustering method does not accommodate studying further this axiomatic assumption, so the questions of whether and to what extent the non-considered within-subset similarities are utilizable for the task of constructing context-dependent correspondences remain open. In addition, the intermediate stage of calculating pairwise similarity implies yet another source for loss of information, which could have contributed to the revealed correspondences.

The cross partition data clustering method, introduced in Chapter 5, extends coupled clustering along several aspects: it enables the identification of correspondences across more than two pre-divided subsets, and it produces probabilistic rather than deterministic clustering output. It also saves the intermediating similarity calculation stage, as it relies on vectorial (probabilistic) representation of the data, which is the original format of the data in many cases. Like coupled clustering, the cross-partition clustering method follows two guiding principles. The first of which is the underlying idea of probabilistic centroid-based clustering (Subsection 2.1.4.6):

> *Clusters are formed around feature-distribution based centroids.*

(The centroids are averaged over individual distributions of members in proportion to their membership probability, and thus expected to approximate the feature distribution for the cluster elements). The other principle is the main novel idea in the cross-partition method:

> *The formed associations between the clusters and the feature distributions characterizing them should maintain independence from the given pre-partition to subsets.*

In order to illustrate the kind of impact this principle has on outcome resulting from the first principle (that is, standard probabilistic clustering), assume that some features distinguish well between groups of elements within one of several pre-determined subsets, while having no discriminative value within other subsets. Such features are *not* expected to direct the formation of cross-partition clusters, as they would push toward clusters made of elements restricted to one subset. Rather, features that push towards inclusion of members from all subsets in some cluster are expected to guide the formation of clusters, even if overall they are not as salient.

This direction is analogous to the second principle guiding the coupled clustering cost term. Both give rise to a geometrical mean term. In the coupled clustering case, the scheme involving geometrical mean has been justified by the intuitive direction of keeping both cluster parts of a considerable size. However, a restrictive working assumption such as the one taken by the coupled clustering method (restricting attention to between subset similarities) is not present in the cross partition framework. As a rough equivalent to our coupled clustering restriction, we mention the use of the maximum entropy principle, which is applied to highlight the constraints posed by the two principles above. Rather than posing some initial guess regarding where to look for the desired information, the maximum entropy principle enforces the assumption that *nothing* is known beyond constraints derived from the stated guiding principles.

There are several technical matters in the cross partition framework that need to be studied further. For example: clarifying the role of the external parameters (η and β) and their interplay and understanding how the number of pre-given subsets ($|W|$) – especially when this number is large – affects the behavior of the algorithm.

An optional direction, where the CP method can be practically applied is identifying repeating themes, or "roles", in topical news articles (which is directly related to the task of template induction for *information extraction*). We did a preliminary investigation in this direction at earlier phases of the research (Marx, Dagan & Shamir, 2002). The news articles that we examined were focused on the topic of terrorist attacks. In this domain, the target roles – the organization that carried out an attack, the location, the weapon used and so on – are typically addressed by different terms in each article. We applied a method that clustered together terms associated with each different role. Thus, each of the generated clusters reveals a correspondence across the given articles, which may underlie a slot in an information extraction template.

A related direction currently being implemented is extending the cross-partition framework to semi-supervised learning. As we indicated (Subsection 2.1.5.3), several recent works proposed to constrain (or "to seed") data clustering, e.g., by pre-specifying lists of element pairs that must, or must not, share a cluster. This idea can be adapted to the information distortion and information bottleneck methods, as well: if assignment probabilities of some data elements are pre-specified (or constrained in other ways), a straightforward modification of the algorithm would minimize a cost term just like the original methods do[1]. By the same token, also the cross partition framework can enable pre-specifying, or constraining, assignments of some of the elements. In an ongoing project, we study a setting where the assignments of all elements of one of the pre-given subsets are known, so this subset forms a training set, while elements of the other subset are assigned to clusters as in the original CP method. The idea mentioned above of applying the maximum entropy principle to highlight these additional constraints through separate iterative update steps is incorporated as well in the same project.

[1] This can be verified with slight modification of lemmas 5.1 and 5.3; specifically, the sums in Eqs. (5.8) and (5.11) should be modified to reflect the constrained assignments.

In this work, we have approached a task related with abstract cognitive functions – construction of correspondences and analogies – through a simple extension of the elementary data-clustering setting. We see our success in coping with a seemingly complicated task by means of a relatively simple setting an appealing aspect of our work. There are additional unsupervised tasks, however, which aim at constructs more complex than standard clustering, for example Bayesian nets, or graphical models. As mentioned, the information bottleneck method described in Section 5.2 has already been generalized to producing more complex types of constructs (*multivariate information bottleneck*; Friedman et al., 2002). Extending the cross partition clustering method along the same direction would form a natural and interesting continuation of the current work, which might lead to more insights regarding analogy making and related tasks.

This work provides an original perspective on the study of analogies. Analogy making is one of those slippery tasks with no consensual pre-given definition or characteristics, but very central to intelligence and creativity. Each one of the existing approaches to analogy making indicates different aspects of analogy as the essential ones (as exemplified in Section 2.2, and discussed a bit further below). In fact, there is a deep disagreement with regard to what are the considerations that underlie analogies in practice (see for example Ch. 4 in Hofstadter et al., 1995). Suggesting a computational framework applicable to this notion, as we have attempted to do here, has been a fascinating challenge, even though it is clear that such suggestion is not going to achieve consensus among researchers in the field.

Our approach to analogy making relies on word co-occurrence distribution in the given data, rather than on hand-written rules or pre-coded data representation of the type used by some previous studies. This approach thus bridges between cognitive motivations and observations regarding analogy making and the familiar vector model, which has been extensively used for practical tasks such as similarity assessment and classification. The data clustering task on which our methods elaborate can be seen as a basic "cognitive" tool for concept discovery. Our work takes this general tool and adapts it to discovery of concepts that form an analogy or other non-trivial context-dependent correspondence.

The setting underlying our computational approach is considerably different than previous views of analogy making. As noted above, the two methods that we introduced ascribe the correspondence being formed to interplay between two different principles. In coupled clustering, these are shared pairwise similarity (across subsets) and simultaneous prominence of the formed cluster in both subsets. In the cross partition method, the underlying factors are communal feature distribution patterns and their independence on the pre-partition variable. To the best of our knowledge, the cross partition framework is the first to characterize correspondence formed across several subsets in terms of statistical independence.

Previous works have considered other issues as central to analogy making. Analogy is ordinarily conceived as a means for problem solving ("analogical reasoning"), an aspect on which we have not focused here. The mapping of relational structure is a crucial conception at the core of the structure mapping theory (Gentner, 1983), but our method does not elaborate on this aspect as well. Further developments of our framework might aim at capturing and emphasizing relational structure.

Several other works also considered the retrieval problem: identifying the optimal object, which would allow the construction of an analogy to a given target object (Forbus, Gentner and Law, 1995). Our work does not address this point, as well: we examine two or more given element subsets, without accounting for how these subsets, or the systems they represent, have been chosen at the first place. The approach that we have introduced, however, is formulated in a general enough manner to allow the incorporation of aspects such as the above ones. For instance, in order to compare the quality of several candidate analogies, we might use cost-related criteria (in coupled clustering), or examine the dynamics of the algorithm (in cross partition clustering; e.g., assessing the quality of candidates according to the β value required for producing a fixed number of clusters).

There is a notable aspect that has been raised by other authors, particularly Hofstadter et al. (1995), which our approach seems to address in some sense: the emergent and fluid nature of the formed solution, which has to do with mental processes of creation and discovery. In some resemblance to the Copycat program (Subsection 2.2.2), our clustering mechanisms are based on aggregation of local changes that gradually evolve to a global solution of the problem at hand, while a temperature variable gradually introduces a more deterministic configuration. Further, our approach allows the formed solution to depend greatly on the context. When a particular subset is matched with different subsets, different themes might be revealed and mapped onto one another, in distinction, for example, from an approach that first clusters each subset independently and then map the independently clustered subsets onto one another. In this respect, as well, our approach accords with Hofstadter et al.'s view regarding the context dependent nature of analogies.

The computational mechanisms that we employ are essentially simpler than the ones suggested by Hofstadter et al. (refer to Subsection 2.2.2.2) and hence they are more liable to inspection and analysis. Hofstadter et al. advocate restricting the scope of investigation to artificial toy problems, allowing "looking at a problem together with its 'hallo' of variant problems" (Hofstadter et al., 1995). We have started with an approach that is inherently simpler. Thus, our approach might capture the analogy making problem only partially (though having potential to incorporate more aspects later on).

Particularly, our methods at their current stage might lack some subtleties addressed within the Copycat program. On the other hand, the principled computational machinery that we suggest allows cleaner demonstration of the impact of systematic manipulations on the input (see description of our experiments with synthetic data, Sections 4.3 and 5.4.1). Yet, without getting to the complex issue of how to evaluate and compare analogies, we think that our method captures something of their emergent and fluid nature. And above all, the most notable advantage we recognize in comparison to previous methods pertaining to analogy making is the immediate applicability of our methods to real-world problems and data.

In this work we have demonstrated our approach mainly on textual data. With no prior specialization or training in the study of religions, our program was able to identify analogous factors shared by several religions in varying levels of resolution: “spiritual” versus “establishment” dimensions in a coarse view and aspects such as “sacred writings”, “rite and festivals” and “sin and suffering” in a more detailed level. These findings are in apparent agreement with previous specialized comparative religion studies that are based on a systematic comparable approach. For the purpose of systematic evaluation, we have measured the overlap between our outcome and religion-related term clusters provided by experts and found their match very close to the level of agreement between experts.

Co-occurrence data and, more generally, data in vectorial representation are very common in many fields: artificial vision, biology, psychology and competitive intelligence, to mention just a few. As the formulation of the methods introduced in this work does not depend on any specific application, we hope they will be applied in the future to a large variety of problems and domains.