Chapter 1: Introduction

Identification of analogies and correspondences has consistently enlightened various fields of knowledge and scholarship. Historical situations and events, for instance, provide a rich field for the construction of analogies. In a unique enterprise, which dates back to Plutarch's "Parallel Lives", each member in a list of Greek public figures is paired with a Roman counterpart, whose position, actions and life events match in an illuminative manner.[1] This classical piece demonstrates one of the fascinating aspects of human intelligence: acquiring knowledge in a context of a particular object – be it a person, an event or a whole domain – allows applying this knowledge and enriching it in a related but different context. To this end, it is required to identify those relevant aspects that correspond to each other across the objects being studied. In other words, utilizing and enriching knowledge along different contexts is grounded on the ability to perceive some compound relation, an equivalence or analogy, between the contexts.

As the ability to perceive correspondence is fundamental to human intelligence, it is interesting to study computational mechanisms underlying it. Along with the theoretical merit, computational methods enabling the automated detection of analogies might turn highly practical in the current information overload era. Consider, for example, the following two text fragments extracted from a pair of 1986 Reuters' news-articles:

1. LOS ANGELES, March 13 – *Computer Memories Inc.* ... agreed to **acquire** *Hemdale Film Corp.* ... That company's **owner**, *John Daly*, would then ***become*** **chief executive officer** of the combined company...

2. NEW YORK, March 25 – *Messidor Ltd* said it signed a letter of intent to **acquire** 100 pct of the outstanding shares of *Triton Beleggineng Nederland B.V.* ... If approved, the **president** of *Triton*, *Hendrik Bokma*, will be ***nominated*** as **chairman** of the combined company. ...

The similarity between the above fragments is apparent: they both deal with the intention of a company to acquire another company. Typically, computational methods for assessing similarity between documents rely on the proportion of shared terms or keywords. In our example the word '*acquire*' appears in both articles, so keyword-based methods would count it as a positive evidence

[1]Plutarch's "Lives" can be browsed at **http://classics.mit.edu/Browse/browse-Plutarch.html**.

for evaluating the text fragments as similar to each other. More sophisticated methods (e.g. *Latent Semantic Indexing*; Deerwester et al., 1990) incorporate term-similarity models that may take into account correspondence of different terms that resemble in their meaning. Thus, corresponding terms such as '*owner*' — '*president*' and '*chief executive officer*' — '*chairman*' may contribute to the unified value of evaluated similarity.

Now, consider yet another pair of terms from the above fragments: '*become*' — '*nominated*'. These terms probably share only a moderate degree of similarity in general, but a human reader will find that they meaningfully correspond to one another in this particular context. Identification of this context-dependent equivalence also enables a reader to perceive that *John Daly* and *Hendrik Bokma* – names that the reader is likely not to encounter before – play in the above texts an analogous part of being appointed to a managerial position. Existing similarity assessment methods do not consider such analogies and do not provide means for pointing them out.

With the goal of spotting context dependent correspondences such as the ones demonstrated above in mind, the present work addresses questions situated one-step ahead of the traditional similarity assessment task: given objects – fragments of news articles in this case – that are already assumed to be similar, *how* they are related to each other? How to identify those aspects of similarity that would facilitate knowledge relevant to both analogized objects?

The research in cognitive science has acknowledged the role of analogy in human thinking. Several works on analogy making have suggested computational mechanisms for constructing detailed mapping that connects corresponding ingredients across a given pair of systems between which analogy is being drawn. According to the *structure mapping* theory (Gentner, 1983), the ingredients of two analogized systems are not expected to share similar individual features, but rather the relations among the ingredients within each system should resemble each other. Another approach (Hofstadter et al., 1995) emphasizes the manner in which features of the analogized objects are perceived in light of the context of aligning them one against the other. Representations that are suitable for mutual mapping are dynamically formed in interaction with the perceived relevance of the features to the correspondence being established. In the next chapter (second part), we review in greater detail these two approaches to analogy making.

The present work is inspired by the direction of constructing a correspondence map, which is grounded on cognitive considerations. We have not found, however, the computational mechanisms employed by cognitive studies directly applicable to real world problems of the type we are interested in, such as identification of corresponding themes in un-annotated texts. One of the above methods (Hofstadter et al., 1995), for instance, has been intrinsically designed for a specific toy problem. This

is justified by the authors' claim that studying toy problems is the best strategy for progress in the field. In their view, the current state of our understanding does not allow more realistic models of analogy making. The other approach (Gentner, 1983) is supposed to be applicable to real world problems of general nature, but it represents information about such problems through pre-coded relational representation. It is not clear if and how information embodied in readily available real world data – free text, for example – can be transformed automatically into this type of relational representation.

Eventually, we have coped with the task of identifying context-dependent correspondences across real-world datasets through adapting up to date computational learning methods. Our work thus bridges between cognitive observations regarding analogy making, which inspired our work but provided no concrete utilizable computational recipes, with techniques that have been proven efficient in processing real world data.

Our strategy is unsupervised: we seek to identify patterns or regularities in the given data without the presence of any examples of suitable correspondences. This approach is practically reasonable, as un-annotated data is freely available in many domains and a newly introduced task would require, on the other hand, a considerable amount of work in preparing training data. Further, it is not clear whether the same factors that underlie a particular correspondence would work in other examples. The use of an unsupervised approach thus makes sense also because it might facilitate something of the non-repeating creative nature of the task.

More specifically, our approach extends recent works on data clustering. *Standard data clustering* methods (a term that we shall used interchangeably with *single-set clustering* to distinguish it from our original elaborations) impose structure of the most elementary form on unstructured data by partitioning the given set of data elements into disjoint clusters. In the next chapter (first part), we refer to the data clustering problem in detail and describe methods addressing it.

In our setting, the given element set is pre-divided to several subsets, and our goal is not just to find the immediate internal structure of each of these subsets but rather to find structure that reveals correspondence between them. In particular, we would like to ignore and, in the more interesting cases, to mask out actively internal structures that are irrelevant to the cross-subset correspondence. To that end, our approach extends the standard clustering task by producing clusters that contain elements of both subsets between which we seek to identify correspondence. Each cluster would thus designate concrete links across ingredients or aspects of systems between which an analogy is drawn. Illustratively, a standard clustering method might cluster together names of employees working for different firms in different clusters. Assuming the employer based partition is pre-given, we would

expect our approach to produce clusters that capture, say, corresponding functions: the management teams of all firms would be clustered together, the same for the sales teams and so on (as opposed to clusters that coincide with firm-specific units). We present this conception in more detail, exemplify it and describe how performance on this task will be evaluated in Chapter 3.

The first method that we introduce, termed *coupled clustering*, is designed for a dataset pre-divided to two disjoint subsets (each associated with one of the analogized systems). We study the problem of partitioning the two subsets into corresponding sub-clusters, so that every such sub-cluster is matched with a counterpart in the other subset. This target is accomplished through elaboration on an axiomatic cost-based framework of pairwise (i.e., proximity based) data clustering (Puzicha, Hofmann & Buhmann, 2000). Puzicha et al.'s original framework is reviewed in Chapter 4, followed by several alternative extensions aiming at our task. The various extensions are tested and evaluated on synthetic and textual data.

In chapter 5, we introduce another method, *cross-partition clustering*, modifying and generalizing 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) partitioning rather than deterministic one and it allows revealing correspondences across more than two subsets. The cross-partition clustering method that we have developed 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. We review in detail these methods before introducing our original elaborations. The cross partition method is tested, as well, on synthetic and textual data, demonstrating noticeable improvement relatively to the coupled clustering results.

Cross partition clustering is a newly defined computational task of general purpose. Potentially, correspondences of the type we study could be drawn across real world composite objects within unrestricted variety of domains. One might be interested, for instance, in identifying corresponding objects in different images. Further examples are discovery of corresponding biological and psychological phenomena typical to different populations, discovery of corresponding business and moves in competing commercial firms (*competitive intelligence*) and so on. The methods introduced in this work are developed using general formulation that would allow adapting them to any application such as the ones mentioned, given data in standard format: similarities between pairs of data elements (coupled clustering) or probabilistic vectorial representation (cross partition clustering).

In accordance with our concrete illustrative examples, the focus of the experimental part of this work is on textual data, specifically, identifying corresponding sub-topics across related, but distinct,

domains (rather than short articles or text fragments as in the previous examples). Each domain is represented by a set of keywords extracted from a corpus of texts discussing it. A keyword is characterized by a vector of its co-occurrences with other words in the corpus. Such co-occurrence based representation, which utilizes the fact that similar words are used in similar lexical contexts, is commonly used for tasks such as estimation of similarity between words and documents. In this work, we utilize the correspondences in context captured by the co-occurrence statistics in order to identify correspondences between groups of words.

More concretely, in Chapter 4, we demonstrate how the coupled clustering method performs on identifying correspondences between conflicts of different nature that were discussed extensively in the new articles. The main body of experimental work, in both Chapters 4 and 5, is concentrated on identifying correspondence between different religions: Buddhism, Christianity, Hinduism, Islam and Judaism. Each of these religions is represented by a collection of texts discussing it. The results (due to both coupled clustering and cross partition clustering) are systematically evaluated against clusters manually produced by experts in the comparative study of religions. Some of our results, particularly those in Chapter 5, reveal fundamental themes common to all religions, which can also be traced in comparative studies on religious.

The computational methods introduced in this work provide novel perspective into the essence of analogies and deep semantic correspondences (as opposed to immediate correspondence, based on superficial appearance). This and further aspects and potential elaborations of our work are discussed in the concluding discussion chapter.