Coupled Clustering: a Method for Detecting Structural Correspondence

Zvika Marx MARXZV@CS.BIU.AC.IL
The Interdisciplinary Center for Neural Computation, The Hebrew University of Jerusalem and
Mathematics and Computer Science Department, Bar-Ilan University, Ramat-Gan 52900, ISRAEL

Ido Dagan IDO.DAGAN@FOCUSENGINE.COM
Mathematics and Computer Science Department, Bar-Ilan University, Ramat-Gan 52900, ISRAEL

Joachim Buhmann JB@INFORMATIK.UNI-BONN.DE
Institut für Informatik III, University of Bonn, Römerstr. 164, D-53117 Bonn, GERMANY

Abstract

This paper proposes a new paradigm and computational framework for identification of correspondences between sub-structures of distinct composite systems. For this, we define and investigate a variant of traditional data clustering, termed *coupled clustering*, which simultaneously identifies corresponding clusters within two data sets. The presented method is demonstrated and evaluated for detecting topical correspondences in textual corpora.

1. Introduction

Numerous studies within (unsupervised) computational learning address two extensive, yet closely related, tasks: assessment of similarity and detection of structure. A common paradigm, for identification and measurement of similarity, associates objects with sets of attributes. A similarity (or distance) value is calculated for each pair of objects, based on these attribute sets. In this scheme, a single value categorically represents the similarity of any pair of compared objects. This approach is useful for various applications: data mining (Das, Mannila, & Ronkainen, 1998), image retrieval (Ortega et al, 1998), syntactic relations approximation (Dagan, Marcus & Markovich, 1995), to mention just few. However, one might wonder how faithfully a single value could represent the whole richness and subtlety of what might be conceived as "similar", for example, the similarity reflected by analogies and metaphors.

Studies concerning structure detection include diversified methods and applications, as well. For example: embedding of data in low dimensional spaces (Deerwester et al., 1990) defining and detecting various equivalency types in data (Batagelj & Ferligoj, 2000), uncovering inherent schemas in semi-structured hierarchical data sources such as the world-wide-web (Kleinberg, 1999; Nestorov et al., 1997). An essential tool for structure detection, which is used in this work, is provided by data clustering methods, revealing the structure emanating from mutual relations of elements within a given system. Such relations might emerge from co-occurrences with a pre-specified set of attributes (Pereira, Tishby & Lee, 1993; Slonim & Tishby, 2000; Dhillon & Modha, 2001), or be represented by pre-calculated values (*pairwise clustering*, Puzicha, Hofmann & Buhmann, 2000) etc.

The current work presents a new approach for analyzing similarity of composite objects: rather than summarizing similarity by one value, we aim to reveal and map corresponding structure. For this purpose, pairwise similarities of data elements (attributes) composing the compared objects are used.

To illustrate our approach, consider a pair of articles addressing two conflicts of distinct types: the Middle East conflict and the dispute over copyright of music and other sorts of media (the "Napster case"). The sample keywords from both texts, listed in Figure 1, can be regarded as the basic elements, or attributes, representing the two domains. As apparent from the presented samples, a naive similarity measure, based on common keyword count, would provide little or no information. This observation is aligned with findings of several previous works: Das et al. (1998) show that similarity assessment, relying solely on the attributes of the examined object pair, can be improved (in terms of retrieval results) by an "external probe", incorporating also additional attributes. For example: two products are similar if similar customers buy them. Likewise, calculation of an overall similarity value could aggregate previously-observed similarities of related term pairs: **delegation**–**committee**, **diplomat**–**lawyer**, **attack**–**infringement**, **security**–**copyright** and so on.

Nevertheless, conceiving any single-valued outcome as a potential oversimplification, our concern is to specify

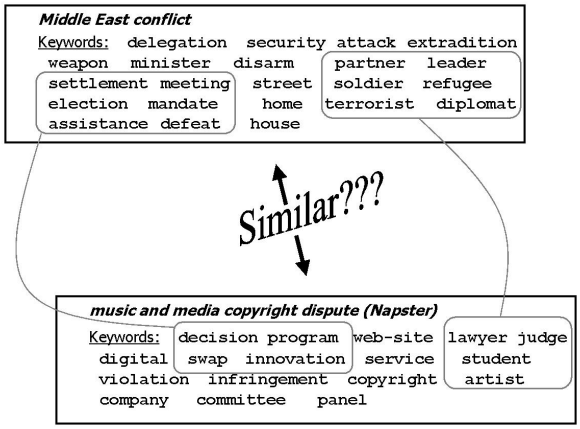

Figure 1. Keyword samples from news articles collections regarding two conflicts. Examples of coupled clusters, i.e. corresponding keyword subsets, are marked with dotted contours.

correspondences of analogous subsets, emerging from elementary relations within the two keyword-sets. We seek to recognize pairs of corresponding keyword subsets referring, for example, to participants involved in the dispute (**diplomat**, **terrorist**, **refugee**, **soldier**… vs. **lawyer**, **judge**, **student**, **artist**…), activities (**election**, **settlement**, **assistance**… vs. **innovation**, **swap**…), assets in dispute (**home**, **house**, **street**, vs. **copyright**, **service**, **web-site**) and so on. We call each such pair of matched subsets a *coupled cluster*. Notice that these correspondences reveal aspects, which are distinctive to the specific context of comparison at hand, but might not hold in other cases (for example, the analogy between **home** and **copyright** as disputed assets).

Searching for corresponding structure, generated by contextually dependent similar components, is inspired by a cognitive theory of analogy – the *Structure Mapping* theory (Gentner, 1983) – viewing analogies as structural equivalence between two systems. One typical example in this research framework illustrates an analogy between an army and an orchestra, comparing soldiers to music players, weapons to musical instruments, and the army commander to the orchestra conductor (Veale, O`Donoghue & Keane, 1999). Computational realizations of this approach were typically applied to knowledge representations that are specially tailored for the algorithm under study (e.g. Falkenhainer, Forrbus & Gentner, 1989). Such mechanisms were subject to criticism regarding their incapability to deduce the context-dependent details that are necessary for analogies from real-world knowledge and representations (Hofstadter et al., 1995).

This paper introduces a conceptual and computational approach to structure mapping, named *coupled clustering*, forming a first step towards automated detection of the context dependent correspondences that are essential for structure based analogies. Criticism, regarding the competence of the structure-mapping approach in coping with real-world data, is addressed by considering features of similarity that are salient in the context of comparing particular systems to one another. The proposed approach is applicable to domains in which corresponding patterns exist and relevant data is available, for example, data mining tasks such as comparison of company profiles (e.g. in *competitive intelligence*). It could be used in other domains, apart from textual data, as well, e.g. for identifying corresponding objects in images.

In earlier work, we have introduced a preliminary version that has included examples of correspondences identified in individual news articles (Marx, Dagan & Shamir, 1999). In the current paper, we propose suitable algorithmic setting for our approach, based on adaptation of relevant aspects from a newly introduced theoretical pairwise clustering framework by Puzicha et al. (2000). Next, we investigate and compare optional versions of this adaptation, through experimentation with synthetic data. Then, using the version that was found best in the synthetic experiments, we apply and evaluate our method through detecting correspondences within textual corpora. We conclude with some directions for future work.

2. Setting and Algorithmic Framework

Our method takes two separate data sets as its input. These sets are assumed to contain the elements (attributes) of the distinct systems for which corresponding structure is to be identified. For example, in the case of textual data these can be keyword sets extracted from two news archives referring to the examined topics. As in the standard task of data clustering, each one of these data sets is partitioned by our method into disjoint subsets. Unlike the standard clustering case, each such subset is coupled to a corresponding subset of the other data set. Each such pair of coupled subsets can be viewed as a unified coupled cluster, containing elements of the two data sets (see Figure 2). Following the identification of word clusters with conceptual categories (cf. Pereira, Tishby and Lee, 1993; Dhillon and Modha, 2001) a list of coupled keyword clusters would reveal correspondence between conceptual entities within distinct content worlds. Alternatively, coupled clusters may contain pixels or contours of corresponding objects in distinct images, etc.

The algorithmic setting, we propose here for coupled clustering, closely follows Puzicha et al. (2000): it is bound to "hard" assignments, i.e. a data element belongs to one and only one subset. The number of coupled clusters is determined in advance. A matrix, whose entries represent pairwise relations between data elements, is introduced as input to the coupled-clustering procedure, in addition to the data elements. We use similarity values where Puzicha et al. use distance (dissimilarity) values, but the transformation is straightforward by interchanging plus and minus signs in subsequent calculations. The effect of these values on the assignment into subsets is realized by a *cost function*, attaching a cost to each conceivable *clustering configuration*. Following Puzicha et

al.'s conclusions, we adopt their cost variant that involves only within-cluster values and weighs each cluster's contribution proportionally to the cluster size.

To this point, the details strictly follow the equivalent details in Puzicha et al. The prominent aspect, in which coupled clustering differs from the general clustering framework, is the set of admissible similarity values to be considered in defining the quality of a clustering configuration, namely those similarity values involved in the cost function. Unlike the case of standard pairwise clustering, where all available similarity values are taken into account, the fundamental premise to coupled clustering states that only *between data-set similarities*, i.e. the similarities of one data set elements to the elements of the other data set, are considered. (These values are represented in Figure 2 as solid and dashed black arrows). Consequently, in our cost formulation, the soft constraint for high average similarity within each cluster is replaced by an equivalent constraint, namely high average of the admissible similarity values within each coupled cluster. This requirement reflects the influence of the context of cross-system comparison on assignments into subsets. Indeed, under this constraint, assignment of a given data element into a subset does not depend on its similarity to members of this subset (solid thin short arrow in Figure 2), but on its similarity to members of the corresponding coupled subset (solid long arrow in Figure 2).

Restricting the set of admissible values to between data-set similarities defines coupled clustering as a novel computational task, different from the well-studied standard data clustering. The input in use – similarity matrix whose rows and columns represent elements of two distinct sets – resembles several previous works on *dyadic data* (Hofmann, Puzicha & Jordan, 1999). Examples for pairs of data sets, which have been investigated in this framework, include verbs and their corresponding transitive subjects (Pereira, Tishby & Lee, 1993) documents and words they contain (Slonim & Tishby, 2000; Deerwester et al. 1990) authoritative and "hub" web pages (Kleinberg, 1999). In these examples, elements of each data set are inter-related with elements of the other set through co-occurrence, containment, web reference relation etc. In distinction from the dyadic setting, coupled clustering is designed for cases in which elements of the examined sets do not share such immediate common relations. Rather, the cross-set similarities (or other types of internal relations) are supposed to be deduced through indirect inference from another (third) class of attributes presented in both compared systems (see section 4 below). This further motivates restricting the considered input to between data-set similarities: in general, the elements of each data set are expected to share much more common attributes among themselves. Including their similarities, as they are, in the calculation would filter out the sensitivity to context of the between data-set similarities. The questions of whether and how to incorporate within data-set similarities is a subject for future research.

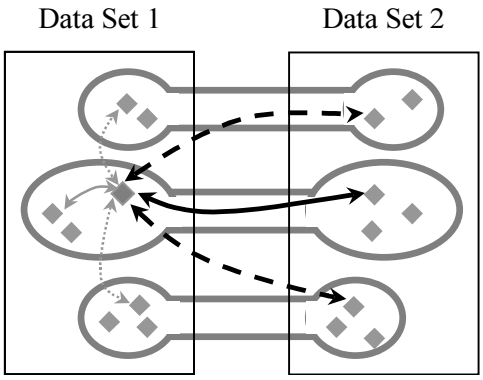

Figure 2. The coupled-clustering setting. The diamond shape objects represent elements of two data sets. Thick closed contours represent coupled clusters boundaries. Long arrows represent the admissible between data set similarities. Short arrows represent the (disregarded) within data set similarities.

We could still use the given asymmetric similarity matrix for distributional clustering (Pereira et al.; Slonim & Tishby, Hofmann et al.) or calculate its principal components (Deerwester et al.; Kleinberg). The results would have explicated, as well, elements of one of the examined data sets, in terms of the elements (attributes) of the other, in low dimensional representation. However, our work aims at reciprocal structure mapping and detection of analogous structure. Therefore, we have chosen to study the (hard) coupled clustering setting, which is directed towards questions such as "what are the context dependently corresponding themes within two domains", rather then "what are the themes and relations emerging in one domain in terms of the other".

In order to search for the best clustering configuration, i.e. the configuration with the lowest cost, we have implemented the *Gibbs Sampler* algorithm for the coupled clustering setting. In brief, this algorithm starts with random assignments and then iterates repeatedly through all data elements and probabilistically reassigns each one of them in its turn according to probability governed by the expected cost change. Specifically, the probability $p(c)$ to reassign a given element into a cluster c is proportional to the *logistic function* $1 / (1 + e^{-\beta \Delta c})$, where Δc is the cost difference obtained from swapping the element's assignment to cluster c and β is an "inverse temperature" parameter. Starting with high temperature, gradual decrease in reassignment randomness is obtained due to gradual "cooling" of the system, i.e. β *is* initialized to a small positive value and gradually increased. This cooling process systematically reduces the probability that the algorithm would get trapped in a local minimum (though obtaining the global minimum is fully guaranteed only under an impracticably slow cooling schedule). The algorithm stops after a few repeated iterations through all data elements, in which no cost reduction has been recorded.

3. Alternative Cost Function Formulations

As the previous section has revealed, the cost is the mechanism through which the admissible similarity values guide identification of qualitative clustering configurations. In this section, alternatives for specific cost formulations are presented and evaluated relative to each other, using results from synthetic data. The cost variant found to be superior in these experiments would be further used, in the following section, for generating coupled clusters of keyword sets.

The coupled clustering setting assumes two given distinct data sets A and B, each containing m and n elements respectively. Also given are similarity values $S = \{s_{ab}\}$ denoting the similarity between every two elements $a \in A$ and $b \in B$ (as explained before, similarities within each data set are disregarded).

Any clustering configuration partitions each one of the data sets into K subsets denoted by A_k, B_k (with sizes m_v, n_v, respectively; $1 \le k \le K$). The same index k, assigned to coupled subsets, denotes that they are both parts of the same unified coupled cluster. In the current setting, the number K of coupled clusters is specified in advance.

A coupled-clustering cost function is expected to assign low costs to configurations in which the similarity values s_{ab} of elements in coupled subsets – $a \in A_k$, $b \in B_k$ – are high on average. (The dual requirement, to assign low costs whenever similarity values s_{ab} of elements of non-coupled subsets – $a \in A_k$, $b \in B_l$, $k \ne l$ – are low, is fulfilled implicitly). In addition, the contribution of large coupled clusters to the cost should be more significant than the contribution of small ones with the same average similarity. Roughly speaking, the reason to this is that we would like to avoid influence of transient or minute components – those that could have been evolved from casual erroneous measurements or during optimization process – and maintain the influence of the stable and macroscopic ones. The manner, in which the cluster's relative contribution is calculated within the cost function, affects the obtained results as is demonstrated below

First, we would like to check the results of applying directly the original cost function by Puzicha, Hofmann & Buhmann (2000) to our restricted set of similarity values. For that, all within-data-set similarities are eliminated, i.e. they are treated as if they are all set to 0. Then the average similarity Avg'_k is calculated as follows:

$$Avg'_k = \frac{\sum_{a \in A_k, b \in B_k} s_{ab}}{(m_k + n_k) \times (m_k + n_k - 1)}$$

For incorporating the size of each cluster, its average similarity is multiplied by its total size: $m_k + n_k$. and the following *standard clustering* cost variant is obtained:

$$(1)\ H_{std} = -\sum_k (m_k + n_k) \times Avg'_k .$$

Alternatively, the expression for average similarity value of the k-th coupled cluster could exclude the zeroed within data-set interactions also from the denominator. In that case, the corresponding size, Avg_k, is given by:

$$Avg_k = \frac{\sum_{a \in A_k, b \in B_k} s_{ab}}{m_k \times n_k}$$

Consequently, a similar *additive* cost variant results:

$$(2)\ H_{add} = -\sum_k (m_k + n_k) \times Avg_k .$$

This cost function obeys all the desired criteria explicated by Puzicha, Hoffman & Buhmann (2000), including *shift invariance* and *strong robustness*. Shift invariance ensures that the cost change, induced by addition of a constant to each one of the admissible similarity values, depends only on the data itself and not on any property of the particular configuration. (Note that in (1) above, adding a constant to all similarities would result in violating the restriction of within data similarities to 0, posed on the standard setting, while in (2) they are inherently disregarded). From here, robustness in the strong sense could be proved, namely the effect of the similarity values concerning a single point is vanished while the data size tends to infinity, or in other words, tolerance to casual erroneous measurements.

A third alternative to weigh coupled clusters with accordance to their relative size is to multiply the average similarity of each cluster by the geometrical mean of the corresponding coupled subset sizes: $(m_k \times n_k)^{1/2}$. The following *multiplicative* cost variant is obtained:

$$(3)\ H_{mult} = -\sum_k (m_k \times n_k)^{1/2} \times Avg_k .$$

This cost function is not strictly shift-invariant. However, it turns to obey a weaker criterion – *shift sub-invariance*: there is a function of the data alone that bounds the change induced by addition of a constant to all similarity values. (The bound is met in clustering configurations in which the global data set proportion is maintained by all coupled sizes, that is $m / n = m_k / n_k$ for each k). This criterion is also sufficient to prove robustness in the strong sense. Note that this variant would yield the same outcome as the *standard clustering* variant H_{std} for configurations in which the global data set proportion is maintained by all coupled sizes ($m / n = m_k / n_k$ for each k).

To investigate and compare the performance of these three cost variants we have conducted a set of experiments on synthetic data. The data for each experiment was a similarity matrix of size $m \times n$, where $m = n = 32$. Each data set was assigned into one of 4 subsets. Each admissible similarity value s_{ab} was set to $(1-x)\delta_{k(a)k(b)} + xr_{ab}$, where $\delta_{k(a)k(b)}$ equals 1 if a and b are in the same coupled cluster and 0 otherwise, $0 \le r_{ab} \le 1$ is a randomly sampled number, and the proportion parameter x varies between 0 to 1.

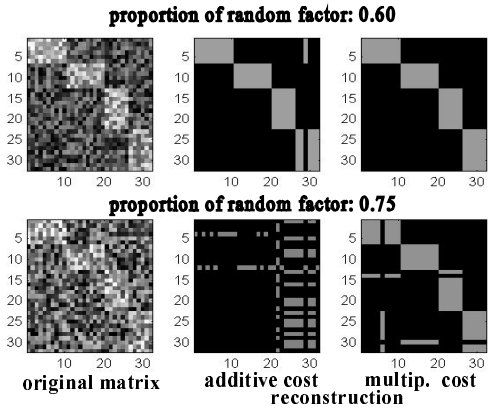

Figure 3. Reconstruction of synthetic noisy coupled-clustering configurations. Lines and columns of the plotted matrices correspond to two distinct data sets. In the left-hand side matrices (original similarity values), the gray level of each pixel corresponds to the similarity value of the appropriate data elements (white – 1; black – 0). In the reconstructed data, gray level correspond to average similarity within reconstructed cluster. The multiplicative cost function seems to relatively maintain its accuracy as the proportion of the random factor grows.

Another varying factor that we have manipulated was the sizes of coupled subsets. We have run 4 sets of experiments in which the corresponding coupled sizes m_k and n_k were set to 8 and 8, 10 and 6, 12 and 4, 14 and 2 respectively. Figure 3 shows examples from the 10-6 experiment set with two distinct levels of randomness. It demonstrates that for this proportion of coupled sizes the multiplicative variant H_{mult} tolerates to noise better than the additive variant H_{add}. We have measured the quality of each reconstructed configuration, in these experiments, by *accuracy*, which is the proportion of data elements that were assigned by the reconstruction into the appropriate subset. Since in cases of poor reconstruction it is not clear which reconstructed subset corresponds to which original subset, the best result obtained by permuting the reconstructed subsets over original subsets was considered.

From the accuracy statistics over the whole range of experiments, displayed in Figure 4, we draw the following observations: The restricted standard clustering H_{std} function results are always worse than the performance of the multiplicative function H_{mult}. In all configurations, there is a range, tending to the left side of the curve, in which the multiplicative function H_{mult} performs better than the additive function H_{add}. This range is almost unnoticeable when the coupled subset sizes are sharply imbalanced and it becomes more prominent as the coupled sizes tend towards maintaining the global proportion. Consequently, it makes sense to use the additive function only if both: (I) there is a good reason to assume that the data contains mostly imbalanced pairs and (II) there is a reason to assume high level of noise. In regular circumstances, where data could be noisy, but there are no indications that the

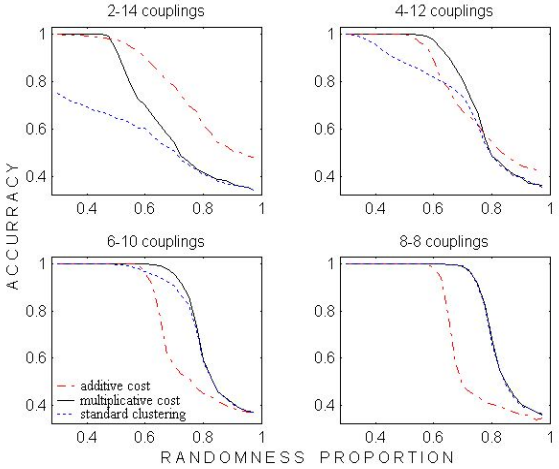

Figure 4. Accuracy as a function of the random factor proportion for different sizes of coupled subsets, obtained through the experiments in reconstruction of synthetic noisy coupled-clustering configurations.

data is inherently imbalanced, the multiplicative function is preferable. Hence, we have used it in the further experiments, described in the sequel.

4. Keyword Coupled Clustering

Textual data benefits, as well, from the preference of balanced coupled clusters, characterizing the H_{mult} cost variant: we have verified that the other variants tend to produce disproportionate clusters, hence we stick to H_{mult} in the rest of our experiments. We have experimented with a variety of textual collections, from which we have extracted sets of key terms to be coupled. For this purpose, the TextAnalyst 2.0 software by MicroSystems Inc. was used (an evaluation copy is available at `http://www.megaputer.com/php/eval.php3`). This software identifies key phrases within each document separately. Since several phrases turn to be inappropriately segmented, and also to avoid rare terms, we have excluded key terms that did not occur in more than one document.

We have extracted key-terms from two collections of news articles (about 30 in each) that were downloaded from the Internet. They were focused on the same topics as in the illustrative example in section 1 above: the Middle East conflict and the music and media copyright dispute. In this experiment, we have used similarity values calculated, by Dekang Lin, from a large collection of news articles, using a co-occurrence-based formulation (Lin, 1998; data is available at `http://armena.cs.ualberta.ca/lindek/downloads/sjmsim.lsp.gz`). The number of coupled clusters was set to 12.

As shown in Table 1, in general, coupled term subsets have fallen, according to our ad-hoc classification, under one of three main types: *Parties and Administration*, *Issues and Resources in Dispute* and *Activities and Procedure*.

Table 1. Conflict coupled clustering data (cluster titles were assigned independently of the coupled clustering procedure).

Middle-East	Music Copyright
Participants & Administration	
establishments	
city state	**company court industry university**
negotiation	
delegation minister	**committee panel**
individuals	
partner refugee soldier terrorist	**student**
professionals	
diplomat leader	**artist judge lawyer**
Issues & Resources in Dispute	
locations	
home house street	**block site**
protection	
housing security	**copyright service**
Activity & Procedure	
resolution	
defeat election mandate meeting	**decision**
activity	
assistance settlement	**innovation program swap**
activity	
disarm extradite extradition face	**use**
confrontation	
attack	**digital infringement label shut violation**
communication	
declare meet	**listen violate**
Poorly-clustered	
low similarity values	
interview peace weapon	**existing found infringe listening medium music song stream worldwide**
no similarity values	
armed diplomatic	

Encouraged by the previous results, we have made thorough experiments concerning the comparison of concepts within different religions. We have downloaded from the Internet three collections of introductory texts regarding the following religions: Buddhism, Christianity and Islam. Key terms were extracted, independently for each religion. The similarity measure we have used, due to Dagan, Marcus & Markovitch (1995), is directed by the proximity of mutual information of both keywords for which similarity is being calculated with each one of their attributes, namely the words co-occurring with them. In other words: the assigned similarity value is high whenever the same attributes predict well occurrences of both keywords. Since two distinct corpora are involved, these attributes are essentially the words commonly used in both corpora (excluding a limited list of stop words). Hence, coupled clustering is directed by the information encapsulated in the words shared by the two particular systems under comparison (note that this case demonstrates further why within corpus similarities, which are significantly higher, could not be straightforwardly incorporated). The results of the Buddhism–Islam key term coupling include, for example, correspondences between scripture-related keyword lists (**Pali**, **Sanskrit**... – **Arabic**, **Hadith**...) as well as lists of terms referring to afterlife and future reward (**pain**, **reborn**... – **judgment**, **paradise**...).

Evaluating results of an unsupervised structure identification procedure is not a straightforward task. To a large extent, this is true even for evaluation of standard similarity assessment and data clustering jobs. It becomes even harder in the case of coupled clustering, where the target structure emerges only whenever two distinct systems are associated with one another. The discipline of comparative religion studies has been supposed to provide some unified grounds for evaluating our results concerning religion key-term coupling. But, it should be noted that even within this discipline, aiming at comparison of distinct systems, there are large deviations between different theories and points of view, as has become apparent from our experimentation. We have asked academics of religion studies to itemize similar aspects of a given pair of religions and to attach to each one of the aspects a list of key-terms from the content-world of each religion. Consequently, a manually performed coupled-clustering configuration is obtained. Looking at the outcome that was provided by two of the experts, it is apparent that there is a wide range of different possible views. For example: one of the lists, regarding *Christianity and Islam*, includes the following cluster titles: “Places and Names”, “Establishments”, “Scripture”, “Beliefs and Ideas ”, “Rituals and Holidays”, “Mysticism”. The items in another list, relating *Buddhism and Christianity*, are titled “Ontological violations of intuition”, “Agency representations”, “Representations concerning religious specialists”, “Representations of causal links between this and another reality”.

Since the provided categories and terms remarkably vary, we have analyzed each expert's contribution separately. Coupled clustering has been performed for the sets of terms given by the experts (in the *Christianity and Islam* coupled clustering, about 30 in each religion; for *Buddhism and Christianity* – about 15 in each religion).

We have analyzed the results in terms of *pairwise recall and precision* of the coupled-clustering configurations induced by our procedure relatively to those given by the experts. Pairwise recall and precision are defined as follows: Denote by *ret* the set of all key-term pairs (each term from another data set) the elements of which belong to subsets that were associated by the expert to the same aspect of similarity. Likewise, denote by *rel* the set of term pairs with both elements in subsets that are coupled by our subset-coupling procedure. Following this notation, the recall is given by $r = (rel \cap ret) \,/\, rel$, and the precision is given by $p = (rel \cap ret) \,/\, ret$. These measures are sensitive to chunks of data that are overlooked by the accuracy measure used for synthetic data, which counts only the members of one key-term list in each subset.

We have not expected our procedure necessarily to fit best expert's data when the number of coupled clusters equal to the number of expert's clusters. For example, one of the experts has taken "Places and Names" to be one unified aspect, but our procedure tends to differentiate it into separate clusters for "Places" and "Names". Another example: our procedure has produced two distinct coupled clusters that are related with the broad notion of "ontological violations of intuition", specified by the other expert as a unified aspect. One of these clusters could be titled "supra-natural entities": **Buddhas**, **Bodhisattvas** and **gods** in Buddhism vs. **Holy Spirit** and **miracles** in Christianity. The other cluster is related to "afterlife": **nirvana**, and **rebirth** in Buddhism vs. **heaven**, **hell** and **resurrection** in Christianity. Yet, it contains terms that have not been initially associated with the broad notion, but are related to the narrower "afterlife" topic: **enlightenment** in Buddhism vs. **angels** and **grace** in Christianity.

Therefore, we have tested our procedure for number of coupled clusters varied form 2 to 10. The recall–precision curves – solid lines in Figure 5 – demonstrate a general tendency of the recall to decrease and the precision do increase as number of coupled clusters grows. (Occasional decreases in precision and increases in recall were filtered out: e.g. the k-th precision value is used also for k+1 clusters whenever the precision calculated for k+1 clusters is lower than for k clusters). Since precision does not improve at all for random coupling, the curves confirm that the results are better than random with relation to the configurations provided by both experts. We conclude that when similarity computation and coupled clustering are performed on terms that are related with a certain unified point of view, the obtained results might reflect the categories of this specific theory. According to further experimentation, as much as the set of key terms is augmented (e.g. to include automatically extracted terms), the match with experts coupling configuration gets worse. Yet, results for automatically extracted terms make sense in their own sake, as demonstrated with the *conflict* data, and there should be found alternative ways to evaluate them. Taking into account the inherent variation in conceivable results of coupled-clustering tasks in general, and within the domain we have checked in particular, we suppose that our procedure demonstrates considerable ability to detect meaningful structure.

We have further compared the coupled clustering with an additional benchmark produced by the following two-stage procedure (dashed curves in Figure 5). At the first stage, each one of the term sets is independently clustered (by the original clustering procedure of Puzicha, Hofmann & Buhmann , 2000). At the second stage, these clusters are paired, so that the best recall, relative to the expert's data, is obtained.

We could see several reasons why this upper bound is not approached in general. One reason is that within data-set similarities, used for clustering, depend, as noted before, on richer co-occurrence information (i.e. much more words co-occurring in common) than the between data-set similarities used for coupled clustering. Another reason is that the separate clustering enjoys an ad-hoc combinatorial advantage. For example, for those parts of the data on which both procedures worked in random, the separate clusters could pick up better pairs of randomly obtained clusters. Finally, our communication with the experts reveals that, to some extent, either consciously or unconsciously, they went through a separate clustering process in order to obtain keyword lists.

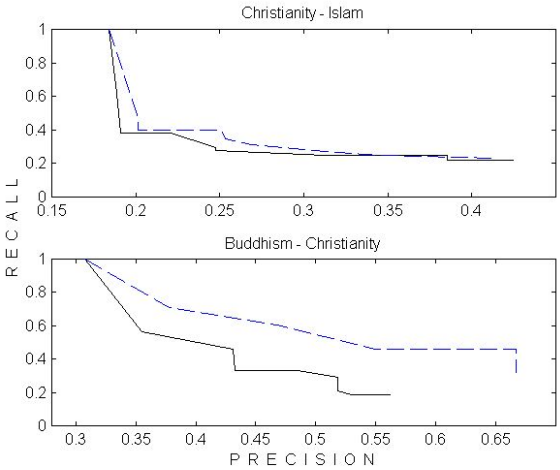

Figure 5. Recall-precision curves for term coupled clustering results, regarding comparison of religions, relative to coupled clusters provided by experts.

This last observation raises questions regarding to how people and theories develop their conceptual categories, in particular theories that are concerned with relative comparison of separate systems. As a starting point for future investigations, we speculate that coupled clustering, relying solely on between data-set relations, could account for theories whose categories relating the systems under study, better than theories identifying matches of previously formed categories. The appropriate way for inclusion of all similarities to improve matching to the current data is left for follow-up research.

5. Conclusions and Future Directions

We have introduced a new framework for unsupervised detection of correspondences between distinct data sets. We have presented and analyzed an algorithmic setting allowing the realization of this framework followed by a preliminary application to real world data. Nevertheless, it is apparent that most of the work is yet to be done.

We have already mentioned the directions of incorporating within system similarities and applying subset coupling to non-textual domains. Another subject for subsequent research is whether the obtained composite picture is reducible to a single value representing an overall level of analogous similarity, though this might largely depend on the specific application. The pairwise setting itself is an arbitrary and tentative choice to start with. We have

mentioned several potential methods (Deerwester et al., 1990; Kleinberg, 1999; Pereira, Tishby & Lee, 1993; Slonim & Tishby, 2000; Hofmann, Puzicha & Jordan, 1999) to be adapted for producing coupled clusters directly from dyadic data generated through an additional set of common attributes, with no mediation of similarity values.

As a yet further guideline, a method to articulate also relations between data elements within each system (emerging, e.g., from syntax) and, furthermore, equivalence between such relations. This would bring coupled clustering closer to the original conception of structure mapping (Gentner, 1983). For this, the notions of *block-modeling* and *structural equivalence*, used in analysis of social networks, provide an interesting direction. Batagelj & Ferligoj (2000) have presented a unified framework relating clustering of relational data and identification of element blocks whose members are structurally equivalent to each other, i.e. have similar relations to other elements. This framework, as most approaches to clustering, refers to a single data set, hence substantially differs from the subset coupling setting. However, it addresses cases in which several relations are considered concurrently, so simultaneous incorporation of our original cross-system approach, with supplementary relations reflecting internal structure, is conceivable.

Acknowledgements

We thank Eli Shamir for ongoing helpful advice. We thank Ilkka Pyysiainen and Eitan Reich for the expert data they have provided, as well as for illuminating discussions.

The first two authors were supported by ISRAEL SCIENCE FOUNDATION founded by The Academy of Sciences and Humanities (grant 574/98-1). This work was also partially supported by a GIF contract I 0403-001 06/95.

References

Batagelj, V. and Ferligoj A. (2000). Clustering relational data.. In W. Gaul, O. Opitz, M. Schader (Eds), *Data Analysis* (pp 3-15), Berlin: Springer.

Dagan, I., Marcus S. and Markovitch S. (1995) Contextual word similarity and estimation from sparse data. *Computer Speech and language*, 9/2, 123-152.

Das G., Mannila H. and Ronkainen P. (1998) Similarity of attributes by external probes. *Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining KDD'98* (pp. 23-29) New York, NY, USA, August 1998. AAAI Press.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis, *Journal of the American Society for Information Science*, 41/6, 391-407.

Dhillon, I. S. and Modha, D. S. (2001). Concept decompositions for large sparse text data using clustering. *Machine Learning*, 42/1, 143-175.

Falkenhainer, B., Forbus, K. & Gentner, D. (1989). The structure mapping engine: Algorithm and example. *Artificial Intelligence*, 41/1, 1-63.

Gentner, D. (1983). Structure-Mapping: A Theoretical Framework for Analogy. *Cognitive Science*, 7/2, 155-170.

Hofmann, T., Puzicha, J. and. Jordan, M. (1999). Learning from Dyadic Data. In*: Advances in Neural Information Processing Systems 11 NIPS*98* (pp. 466-472).

Hofstadter, D. R. and the Fluid Analogies Research Group (1995). *Fluid Concepts and Creative Analogies*. New-York: Basic Books.

Kleinberg, J. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, 46/5, 604-632.

Lin, D. (1999). Automatic Retrieval and Clustering of Similar Words. *Proceedings of the 17th International Conference on Computational Linguistics and the 36th Annual Meeting of the Association for Computational Linguistics* COLING-*ACL '98* (pp. 768-774). Montreal.

Marx, Z., Dagan, I. and Shamir, E. (1999) Detecting Sub-Topic Correspondence through Bipartite Term Clustering. *Proceedings of the ACL-99 Workshop on Unsupervised Learning in Natural Language Processing* (pp. 45-51). College Park, MD.

Nestorov, S., Ullman, J., Wiener, J. and Chawathe, S. (1997). Representative Objects: Concise Representations of Semistructured, Hierarchical Data. *Proceedings of the 13th International Conference on Data Engineering ICDE'97*, (pp. 79-90), Birmingham, U.K..

Ortega, M., Rui, Y., Chakrabarti, K., Mehrotra, S. and Huang, T. (1998). Supporting ranked boolean similarity queries in mars. *IEEE Transactions on Knowledge and Data Engineering*, 10/6, 905-925.

Pereira, F. C. N., Tishby N. Z. and Lee L. J. (1993). Distributional Clustering of English Words. *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics ACL' 93* (pp. 183-190). Columbus, OH.

Slonim, N. and Tishby, N. (2000). Agglomerative information bottleneck. In *Advances in Neural Information Processing Systems 13 NIPS*2000*, (pp. 617-623).

Veale, T., O`Donoghue D. and Keane M. T. (1999). Computability as a Limiting Cognitive Constraint: Complexity Concerns in Metaphor Comprehension. In M. K. Hiraga, C. Sinha and S. Wilcox (Eds) , *Cultural, Psychological and Typological Issues in Cognitive Linguistics*. Amsterdam: John Benjamins.