Chapter 4: Coupled Clustering

The method introduced in this chapter – *coupled clustering* – extends the standard clustering task for a set of data elements pre-divided into two disjoint subsets. As explained in the previous chapter, we study the problem of partitioning in parallel the pair of pre-given subsets to groups of elements, or sub-clusters, each of which is matched with a corresponding sub-cluster in the other subset. A pair of matched sub-clusters forms together a *coupled cluster*.

This chapter relies on a paper introducing the coupled clustering framework (Marx et al., 2002), extending earlier publications (Marx & Dagan, 2001; Marx, Dagan & Buhmann, 2001). In Section 4.1, we review the computational methods that underlie our approach, namely the standard clustering method by Puzicha, Hofmann & Buhmann (2000) and co-occurrence based similarity measures by Lin (1998) and Dagan, Marcus & Markovitch (1995). The coupled clustering method is formally introduced in Section 4.2. Then, we demonstrate our method on synthetic data (Section 4.3) and on a task of detecting equivalencies across distinct textual corpora (Section 4.4). The examined corpora deal with the conflict theme as exemplified in Figure 3.2 and on various religions between which we identify correspondences. The religion data is thoroughly evaluated through comparison of our program's output with keyword classes that were formed manually by experts of comparative studies of religions. Section 4.5 concludes this chapter with further discussion.

4.1 Computational Background

This section reviews the two computational frameworks that form the basis for the coupled-clustering method. The first subsection concentrates on the relevant details of a data-clustering method, by Puzicha, Hofmann & Buhmann (2000), which our algorithm extends for coupled clustering. The next subsection reviews methods for calculating the similarity values used as input for our method.

4.1.1 Cost-based Pairwise Clustering

Puzicha, Hofmann & Buhmann (2000) present, analyze and classify a family of pairwise clustering cost functions. Their framework assumes "hard" assignments: every data element is assigned into one and only one of the clusters. In reviewing their work we use the following notation. A data clustering procedure partitions the elements of a given dataset, $\boldsymbol{X}$, into disjoint element clusters, $c_1, \ldots, c_k$. The number of clusters, k, is pre-determined and specified as an input parameter to the clustering algorithm. A cost criterion guides the search for a suitable clustering configuration. This criterion is realized through a cost function $H(S, \boldsymbol{C})$ taking the following parameters:

(i) $S = \{s_{xx'}\}_{x,x' \in \boldsymbol{X}}$: a collection of pairwise similarity values[1], each of which pertains to a pair of data elements x and x' in $\boldsymbol{X}$.

(ii) $\boldsymbol{C} = (c_1, \ldots, c_k)$: a candidate clustering configuration, specifying assignments of all elements into the disjoint clusters (that is $\bigcup c_j = \boldsymbol{X}$ and $c_j \cap c_{j'} = \phi$ for every $1 \leq j \leq j' \leq k$).

The cost function outputs a numeric cost value for the input clustering-configuration $\boldsymbol{C}$, given the similarity collection S. Thus, various candidate configurations can be compared and the best one, i.e the configuration of lowest cost, is chosen. The main idea underlying clustering criteria is the preference of configurations in which similarity of elements within each cluster is generally high and similarity of elements that are not in the same cluster is correspondingly low. This idea is formalized by Puzicha et al. through the *monotonicity* axiom: in a given clustering configuration, increasing similarity values, pertaining to elements within the same cluster, cannot increase the cost assigned to that configuration. Similarly, increasing the similarity level of elements belonging to distinct clusters cannot improve the cost.

Monotonicity captures the most basic intuitive expectation from pairwise data clustering. By introducing further requirements, Puzicha et al. focus on a more confined family of cost functions. The following requirement focuses attention on functions of relatively simple structure. A cost function H fulfills the *additivity* axiom if it can be presented as the cumulative sum of repeated applications of "local" functions referring individually to each pair of data elements. That is:

$$H(S, \boldsymbol{C}) \;=\; \sum_{x,x' \in \boldsymbol{X}} \psi^{xx'}(x, x', s_{xx'}, \boldsymbol{C}) \,, \qquad (4.1)$$

where $\psi^{xx'}$ depends on the two data elements x and x', their similarity value, $s_{xx'}$, and the whole clustering configuration $\boldsymbol{C}$. An additional axiom, the *permutation invariance* axiom, states that cost should be independent of element and cluster reordering. Combined with the additivity axiom, it implies that a single local function ψ, s.t. $\psi^{xx'} \equiv \psi$ for all $x,x' \in \boldsymbol{X}$, can be assumed.

Two additional invariance requirements aim at stabilizing the cost under simple transformations of the data. First, relative ranking of all clustering configurations should persist under scalar multiplication of the whole similarity ensemble. Assume that all similarity values within a given collection S are multiplied by a positive constant η, and denote the modified collection by ηS. Then, H fulfills the *scale invariance* axiom if for every fixed clustering configuration $\boldsymbol{C}$, the following holds:

---

[1] In their original formulation, Puzicha et al. use distance values (dissimilarities) rather then similarities. Hereinafter, we apply straightforward adaptation to similarity values by adding a minus sign to H. Adhering to the cost minimization principle, this transformation replaces the cost paid for within-cluster dissimilarities with cost saved for within-cluster similarities (alternatively pronounced as "negative cost paid").

$$H(\eta S, \boldsymbol{C}) \;=\; \eta H(S, \boldsymbol{C})\,. \tag{4.2}$$

Likewise, it is desirable to control the effect of an addition of a constant. Assume that a fixed constant Δ is added to all similarity values in a given collection S, and denote the modified collection by $S^{+\Delta}$. Then, H fulfills the *shift invariance* axiom if for every fixed clustering configuration $\boldsymbol{C}$, the following holds:

$$H(S^{+\Delta}, \boldsymbol{C}) \;=\; H(S, \boldsymbol{C}) + \Phi\,, \tag{4.3}$$

where Φ may depend on Δ and on any aspect of the clustered data (typically the data size), but not on the particular configuration $\boldsymbol{C}$.

As the most consequential criterion, to assure that a given cost function is not subject to local slips, Puzicha et al. suggest a criterion for *robustness*. This criterion ensures that whenever the data is large enough, bounded changes in the similarity values regarding one specific element, $x \in \boldsymbol{X}$, would result in limited effect on the cost. Consequently, the cost assigned to any clustering configuration would not be sensitive to a small number of fluctuations in the similarity data. Formally, denote the size of the set of elements $\boldsymbol{X}$ by N and let $S^{x+\Delta}$ be the collection obtained by adding Δ to all similarity values in S pertaining to one particular element, $x \in \boldsymbol{X}$. Then H is robust (in the strong sense) if it fulfills

$$\tfrac{1}{N}\,|\,H(S, \boldsymbol{C}) - H(S^{x+\Delta}, \boldsymbol{C})\,| \;\xrightarrow[N\to\infty]{}\; 0. \tag{4.4}$$

Puzicha et al. show that any cost function satisfying Equations 4.1, 4.2, 4.3 is a linear combination of two factors: a positive component (to be minimized) incorporating averages of distances between elements within the same cluster, and a negative component (to be maximized) incorporating averages of distances between elements from different clusters. It turns out that among those cost functions there is only one function that satisfies the strong robustness criterion of Equation 4.4 in addition to Equations 4.1, 4.2, 4.3. This function, denoted here as H^0, involves only similarity values pertaining to elements within the same cluster (*within-cluster similarities*).

Specifically, H^0 is a weighted sum of average within-cluster similarity. Denote the sizes of the k clusters $c_1, \ldots, c_k$ by $n_1, \ldots, n_k$ respectively. The average within-cluster similarity for the cluster c_j is then

$$Avg_j \;=\; \frac{\sum_{x,x' \in c_j} s_{xx'}}{n_j \times (n_j - 1)}\,. \tag{4.5}$$

H^0 weights the contribution of each cluster to the cost proportionally to the cluster size:

$$H^0 \;=\; -\sum_j n_j Avg_j\,. \tag{4.6}$$

In Section 4.3, we modify H^0 to adapt it for the coupled clustering setting.

4.1.2 Feature-based Similarity Measures

In our calculations, similarity between data elements is assessed through methods that take feature vectors as input and put heavier weights on the more informative features. The information regarding a data element, x, conveyed through a given feature, y, is assessed through the following term:[2]

$$I(x,y) = \log_2^+ \frac{p(x \mid y)}{p(x)} , \tag{4.7}$$

where, p denotes conditional and unconditional occurrence probabilities and the '+' sign indicates that 0 is returned whenever the $\log_2$ function produces negative value. In our experiments x and y are generally words and p is empirical occurrence probability.

Dagan, Marcus & Markovitch (1995) base their similarity measure on the following term:

$$sim_{DMM}(x,x') = \frac{\sum_y \min\{I(x,y), I(x',y)\}}{\sum_y \max\{I(x,y), I(x',y)\}} . \tag{4.8}$$

The similarity value obtained by this measure is higher as the number of highly informative features, providing comparable amount of information for both elements x and x', is larger.

Lin, 1998 incorporates the information term of Equation 4.7, as well, though differently:

$$sim_L(x,x') = \frac{\sum_{\{y/I(x,)>0 \wedge I(x',y)>0\}} (I(x,y) + I(x',y))}{\sum_y (I(x,y) + I(x',y))} . \tag{4.9}$$

Here, the obtained similarity value is higher as the number of features that are somewhat informative for both elements, x_1 and x_2, is larger, and the relative contribution of those is in proportion to the total information all features convey.

Similarly to the cosine measure (see 2.1.3.3), both sim_{DMM} and sim_L measures satisfy: (i) the maximal similarity value, 1, is obtained for element pairs with relation to which every feature is equally informative (including self similarity); and (ii) the minimal similarity value, 0, is obtained whenever every attribute is not informative for either one of the elements. Accordingly, our formulation and the experiments below follow the convention that a zero value denotes no similarity.

In the coupled clustering experiments on textual data that are described later, we use both above similarity measures. We utilize pre-calculated sim_L values for one experiment (Subsection 4.4.1) and we calculate sim_{DMM} values, based on word co-occurrence within our corpora, for another experiment (Subsection 4.4.2).

---

[2] The expectation over the term of Equation 4.7 over co-occurrences of all x's and y's, (with $\log_2$) is the *mutual information* of x and y (Cover and Thomas, 1991, p. 18).

4.2 Algorithmic Framework for Coupled Clustering

In this section, we define the coupled clustering task and introduce an appropriate setting for accomplishing it. We then present alternative cost criteria that can be applied within this setting and describe the search method that we use to identify coupled-clustering configurations of low cost. As we noted in Chapter 3, coupled clustering is the problem of partitioning two data subsets into corresponding sub-clusters, so that every sub-cluster is matched with a counterpart in the other subset. Each pair of matched sub-clusters forms jointly a coupled cluster. As in the standard single set task of data clustering, each coupled cluster consists of elements that are similar to one another and distinct from elements in other clusters. However, this is subject to an additional bias imposed by the requirement to match sub-clusters of each pre-given subsets with those of the other subset.

4.2.1 Directing Clustering through Between-subset Similarities

Coupled clustering divides the two pre-given element subsets, X^1 and X^2, into disjoint sub-clusters $c_1^1, \ldots, c_k^1$ and $c_1^2, \ldots, c_k^2$. Each of these sub-clusters is coupled with a corresponding sub-cluster of the other subset, that is c_j^1 is coupled with c_j^2 for $j = 1 \ldots k$. Every pair of coupled sub-clusters forms a unified coupled-cluster, $c_j = c_j^1 \cup c_j^2$, which contains elements of both pre-given subsets (see Figure 4.1). We approach the coupled clustering problem through a pairwise-similarity-based setting, incorporating the elements of both X^1 and X^2. Our treatment is independent of the method by which similarity values are calculated: feature-based calculations such as those described in Subsection 4.1.2, subjective assessments, or any other method.

The notable feature distinguishing our method from standard pairwise clustering, is the set of similarity values, S, that are considered. A standard pairwise clustering procedure potentially considers the similarity values referring to all pairs of elements within the undivided clustered set. Typically, the only similarity values that are not considered are self-similarities. In the coupled clustering setting, there are two different types of available similarity values. Values of one type denote within-subset similarities (short gray arrows in Figure 4.1). Values of the second type denote similarities of element pairs consisting of one element from each subset (*between-subset similarities*; long black arrows in Figure 4.1). As an initial strategy, to be complied with throughout this chapter, we choose to ignore similarities of the first type altogether and to concentrate solely on between-subset similarities: $S = \{s_{xx'}\}$, where $x \in X^1$ and $x' \in X^2$. Consequently, the assignment of a given data element into a coupled cluster is directly influenced by the most similar elements of the other subset, regardless of its similarity to members of its own subset.

The policy of excluding within-subset similarities captures, according to our conception, the unique context posed by aligning two pre-given subsets representing distinct domains with respect to one

another. Correspondences special to the current comparison, which underlie presumed parallel or analogous structure of the compared systems, are thus likely to be identified abstracted from the distinctive information characterizing each individual system. This fits our key goal of detecting commonalities, while masking out subset-internal structures (see Section 3.1). In this chapter, we do not deal with the questions of whether and how available information regarding within-subset similarities should be incorporated. The next chapter introduces a method that processes the data more comprehensively.

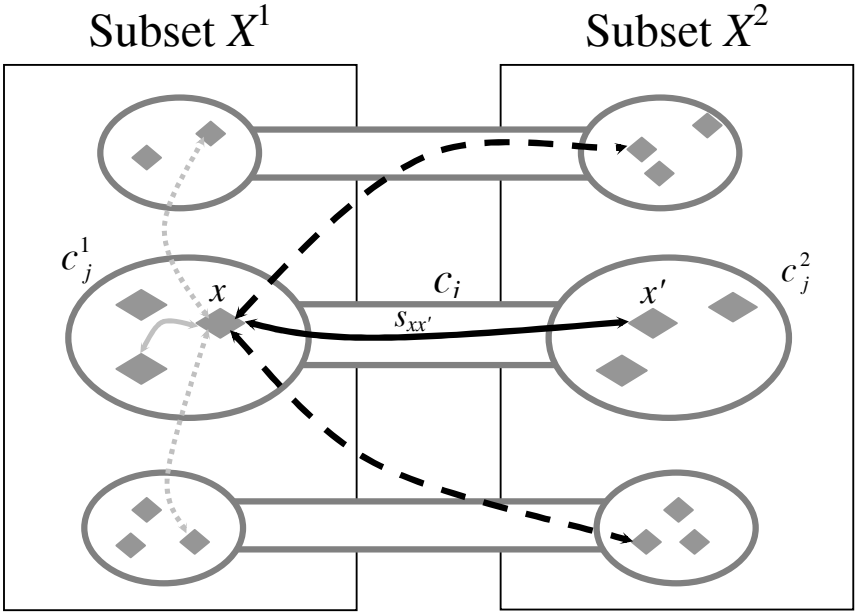

Figure 4.1: The coupled clustering setting. The diamonds represent elements of the pre-given subsets X^1 and X^2. The long black arrows represent the values in use: similarity values pertaining to two elements, one from each subset. The shorter grey arrows stand for the disregarded similarity values within a subset.

4.2.2 Three Alternative Coupled Clustering Cost Functions

Given the setting described above, in order to identify configurations that accomplish the coupled clustering task, our next step is defining a cost function. In formulating it, we closely follow the standard pairwise-clustering framework presented by Puzicha, Hofmann & Buhmann, (2000, see Subection 4.1.1 above). Given a collection of similarity values S pertaining to the members of two pre-given subsets, X^1 and X^2, we formulate an additive cost function, $H(S, \boldsymbol{C})$, which assigns a cost value to any coupled-clustering configuration $\boldsymbol{C}$. Given such a cost function and a search strategy (see 4.2.4 below) our procedure would be able to output a coupled clustering configuration specifying

assignments of the elements into a pre-determined number, k, of coupled clusters. We concentrate on Puzicha et al.'s H^0 cost function (Subection 4.1.1, Eq. 4.6), which is limited to similarity values within each cluster and weighs each cluster's contribution proportionally to its size. Below we present and analyze three alternative cost-functions derived from H^0.

As in clustering in general, the coupled clustering cost function should assign similar elements into the same cluster and dissimilar elements into distinct clusters (as articulated by the monotonicity axiom in Subsection 4.1.1). A coupled-clustering cost function is thus expected to assign low cost to configurations in which the similarity values, $s_{xx'}$, of elements x and x' of coupled subs-clusters, c_j^1 and $c_{j'}^2$, are high on average. (The dual requirement to assign low cost whenever similarity values of elements x and x' from non-coupled sub-clusters c_j^1 and $c_{j'}^2$, $j \neq j'$, are low, is implicitly fulfilled). In addition, we seek to avoid influence of transient or minute components – those that could have been evolved from casual noise or during the optimization process – and maintain the influence of stable larger components. Consequently, the contribution of large coupled clusters to the cost is greater than the contribution of small ones with the same average similarity. This direction is realized in H^0 through weighting each cluster's contribution by its size.

In the coupled-clustering case, one apparent option is to apply straightforwardly the original H^0 cost function to our restricted collection of between-subset similarity values. The average similarity of the coupled cluster $c_j = c_j^1 \cup c_j^2$ is then calculated as

$$Avg'_j = \frac{\sum_{x \in c_j^1, x' \in c_j^2} s_{xx'}}{n_j \times (n_j - 1)}, \tag{4.10}$$

where n_j is the number of elements in c_j (so that Eq. 4.10 differs from the standard average formula, Eq. 4.5, by setting all within-subset similarities to 0). As in H^0 (Eq. 4.6), the average similarity of each cluster is multiplied by the cluster size. Thus, the following cost function, H^1, is obtained:

$$H^1 = -\sum_j n_j \times Avg'_j . \tag{4.11}$$

Alternatively, as we restrict the collection of similarities being considered in our calculations, we might want to take it into account in the averaging scheme as well. The actual number of considered similarities in the restricted collection is, for each j, the product $n_j^1 \times n_j^2$ of the sizes of the two sub-clusters c_j^1 and c_j^2 forming c_j. The following averaging scheme might seem more natural for the coupled clustering setting:

$$Avg''_j = \frac{\sum_{x \in c_j^1, x' \in c_j^2} s_{xx'}}{n_j^1 \times n_j^2}, \tag{4.12}$$

Correspondingly, a second cost variant, H^2, is given:

$$H^2 \;=\; -\sum_j n_j \times Avg''_j \,. \tag{4.13}$$

One factor to which the weighting schemes of H^1 and H^2 does not refer is the inner partition of the coupled clusters. Hence, we suggest yet another alternative that incorporates the proportion between the sizes of the two sub clusters, namely weighing the average similarity each cluster contributes to the cost by the geometrical mean of the corresponding coupled sub-cluster sizes: $\sqrt{n_j^1 \times n_j^2}$. This yields yet another cost function:

$$H^3 \;=\; -\sum_j \sqrt{n_j^1 \times n_j^2} \;\times Avg''_j . \tag{4.14}$$

The weighting factor of H^3 results in penalizing large gaps between the two sizes, n_j^1 and n_j^2, and in preferring balanced configurations, with coupled-cluster inner proportions maintaining the global proportion of the clustered subsets ($n_j^1 / n_j^2 \cong N^1 / N^2$ for each j, where N^1 and N^2 are the sizes of X^1 and X^2, respectively). Later on we refer to the cost function H^2 and H^3, as the "additive" and the "multiplicative" cost functions.

4.2.3 Properties of the Coupled Clustering Cost Functions

Puzicha, Hofmann & Buhmann, (2000) based their characterization of pairwise-clustering cost-functions on some properties and axioms (see Subsection 4.1.1 above). In the previous subsection, we have followed their conclusions in adapting, in three different variants, one function, H^0, that realizes the most favorable properties. It is worthwhile to see if and how these properties are preserved through the adaptation for the coupled clustering setting. As we show below, all the three cost functions that we have derived, H^1, H^2 and H^3, are additive by construction and it immediately follows that they are also scale invariant. They are not, except for H^2, shift-invariant. However, the effect of a constant added to all between-subset similarity values is bounded for H^1 and H^3, as well. Finally, H^1 and H^3 are robust (but not by H^2).

Lemma 4.1: H^1, H^2 and H^3, are additive (Eq. 4.1).

Proof: For each one of the three functions, each element pair with non-zero impact on the cost (i.e., members of the same coupled cluster from different subsets) adds to the cost a component of the form $\psi^{xx'}(x, x', s_{xx'}, \boldsymbol{C})$. This contribution amounts to the average similarity within the cluster to which both elements belong multiplied by a factor depending on this cluster. Specifically, for each pair of elements $x, x' \in c_j$, such that $x \in X^1$ and $x' \in X^2$, we have the following terms:

$$\psi_1^{xx'} = -\frac{s_{xx'}}{n_j - 1} \;,\quad \psi_2^{xx'} = -n_j \frac{s_{xx'}}{n_j^1 \times n_j^2},\quad \psi_3^{xx'} = -\frac{s_{xx'}}{\sqrt{n_j^1 \times n_j^2}}, \tag{4.15}$$

which explicate the non-zero summands forming H^1, H^2 and H^3, respectively. □

Lemma 4.2: H^1, H^2 and H^3, are scale invariant (Eq. 4.2).

Proof: As $\psi_1^{xx'}$, $\psi_2^{xx'}$ and $\psi_3^{xx'}$ of the previous lemma all depend linearly on $s_{xx'}$ (i.e., $\eta\psi_i^{xx'}(x, x', s_{xx'}, \boldsymbol{C}) = \psi_i^{xx'}(x, x', \eta s_{xx'}, \boldsymbol{C})$), it follows that the three cost functions satisfy the scale invariance property. □

Lemma 4.3: H^2 is shift invariant (Eq. 4.3).

Proof: in the j-th cluster there are $n_j^1 \times n_j^2$ cross-subset pairs (x, x'), so that introducing a constant shift to all the considered similarities and summing $\psi_2^{xx'}$ (defined in the proof to Lemma 5.1) over all the relevant pairs within the j-th cluster gives:

$$\sum_{x,x' \in c_j, x \in X^1, x' \in X^2} -n_j \frac{s_{xx'} + \Delta}{n_j^1 \times n_j^2} \quad = \quad \left(\sum_{x,x' \in c_j, x \in X^1, x' \in X^2} -n_j \frac{s_{xx'}}{n_j^1 \times n_j^2} \right) - n_j \Delta . \qquad (4.16)$$

Taking sum over the above cluster-dependent terms of Eq. 4.16 yields $H^2(S^{+\Delta}, \boldsymbol{C}) = H^2(S, \boldsymbol{C}) + N\Delta$, where the term $N\Delta$ depends on the shift constant and on the data (and not on $\boldsymbol{C}$), as required for maintaining the shift invariance property. □

Lemma 4.4: For H^1 and H^3, the effect on the cost value of adding a constant to all between-subset similarities is bounded.

Proof: Both H^1 and H^3 are are non-positive, thus bounded from above by 0.

For H^1, the j-th cluster's contribution to the modified cost resulting from increasing all similarities by a positive Δ is $(n_j^1 \times n_j^2 / (n_j - 1))\Delta \leq \min\{n_j^1, n_j^2\}\Delta \leq (n_j/2)\Delta$. Therefore, $H^1(S^{+\Delta}, \boldsymbol{C}) \geq H^1(S, \boldsymbol{C}) - (N/2)\Delta$.

Similarly, for H^3, the j-th cluster's contribution to the modification in cost value is $\sqrt{n_j^1 \times n_j^2}\,\Delta \leq (n_j/2)\Delta$, so that also for H^3 the following holds: $H^3(S^{+\Delta}, \boldsymbol{C}) \geq H^3(S, \boldsymbol{C}) - (N/2)\Delta$. □

We note that it is possible to modify H^1 and H^3 so to impose the shift-invariance property on them. For that, one can use the derivative of, say, H^3 with respect to Δ, which is the increment for all between-data-set similarity values. This is a linear function so the resulting derivative is $D = \sum_j 1/\sqrt{n_j^1 \times n_j^2}$. Consequently, normalizing H^3 by $1/D$ would result in perfect shift invariance. However, H^3 in its non-normalized form is nearly shift-invariance with regard to configurations for which the clusters approximately maintain the global proportion of the clustered data sets X^1 and X^2, while highly imbalanced configurations are highly penalized. Since our experiments use similarity measures with values between 0 and 1, we stick to the simpler formulation of Eqs. 4.11 and 4.14 above, assuming that the normalized version would behave similarly.

Lemma 4.5: H^1 is robust (Eq. 4.4); H^3 is robust provided the ratio between the sub-cluster sizes of any coupled cluster is kept bounded as the number of elements grows.

Proof: For H^1:

$$\frac{1}{N}\left|H^1(S,\boldsymbol{C})-H^1(S^{x+\Delta},\boldsymbol{C})\right| \;\le\; \frac{1}{N}\max\{n_j^1,n_j^2\}\frac{\Delta}{n_j-1} \;\le\; \frac{\Delta}{N} \xrightarrow[N\to\infty]{} 0\,, \qquad (4.17)$$

where j is the index of the cluster to which x is assigned. Similarly for H^3:

$$\frac{1}{N}\left|H^3(S,\boldsymbol{C})-H^3(S^{x+\Delta},\boldsymbol{C})\right| \qquad (4.18)$$

$$\le\; \frac{1}{N}\max\{n_j^1,n_j^2\}\frac{\Delta}{\sqrt{n_j^1\times n_j^2}} \;=\; \frac{\Delta}{N}\sqrt{\frac{\max\{n_j^1,n_j^2\}}{\min\{n_j^1,n_j^2\}}} \xrightarrow[N\to\infty]{} 0\,,$$

where convergence relies on the assumption regarding the ratio between the sub-clusters. □

Finally, we note that using the coupled sizes geometrical mean as a weighting factor, H^3 tends to escape configurations that match minute sub-clusters with large ones, which are occasionally the consequence of noise in the input data or of fluctuations in the search process. It turns that this property provides H^3 with a notable advantage over H^1 and H^2, as our experiments indeed show (see Sections 4.3 and 4.4).

4.2.4 Optimization Method

In order to find the clustering configuration of minimal cost, we have implemented a stochastic search procedure, namely a variation of the simulated annealing method based on the sampling pattern of the Gibbs sampler algorithm (Geman & Geman, 1984; See also Chapter 2, Subsection 2.1.4.5). Starting with random assignments into clusters, this algorithm iterates repeatedly through all data elements and probabilistically reassigns each one of them in its turn, according to a probability governed by the expected cost change. Suppose that in a given assignment configuration, $\boldsymbol{C}$, the cost difference $\Delta_{j|x,\boldsymbol{C}}$ is obtained by reassigning a given element, x, into the j-th cluster ($\Delta_{j|x,\boldsymbol{C}} = 0$ in case x is already assigned to the j-th cluster). The target cluster, into which the reassignment is actually performed, is selected among all candidates with probability

$$p(j) \;\equiv\; p(j|x,\boldsymbol{C}) \;\prec\; \frac{1}{1+\exp\{-\beta\Delta_{j|x,C}\}} \qquad (4.17)$$

Consequently, the chances of an assignment to take place are higher as the resulting reduction in cost is larger. In distinction from the original simulated annealing algorithm (Kirkpatrick, Gelatt & Vecchi, 1983), assignments that result in increased cost are possible, though with relatively low probability. The β parameter, controlling the randomness level of reassignments, functions as an inverse "computational temperature". Starting at high temperature followed by a progressive cooling

schedule, that is initializing β to a small positive value and gradually increasing it (e.g. repeatedly multiply β by a constant that is slightly greater than one, 1.001 in our experiments), turns most profitable assignments increasingly probable. As the clustering process proceeds, the gradual "cooling" systematically reduces the probability that the algorithm would be trapped in a local minimum (though global minimum is fully guaranteed only under an impracticably slow cooling schedule). The algorithm execution stops after several repeated iterations through all data elements, in which no cost change has been recorded (50 iterations in our experiments).

4.3 Experiments with Synthetic Data

A set of experiments on synthetic data has been conducted for evaluating the performance of our algorithm, making use of the three cost functions introduced in Section 4.3 above. These experiments have measured, under changing noise levels, how well each of the cost functions reconstructs a configuration of pre-determined clusters of various inner proportions.

Each input similarity value (i.e. between-subset similarities) in these experiments incorporates a basic similarity level, dictated by the pre-determined clustering configuration, combined with an added random component introducing noise. The basic similarity values have been generated so that each element is assigned into one of four coupled clusters. Elements in the same cluster share the maximal basic similarity of value 1, while elements in distinct clusters share the minimal basic similarity 0. The noisy component combined with the basic value is a random number between 0 and 1.

In precise terms, the similarity value $s_{xx'}$, of any $x \in X^1$ and $x' \in X^2$ (X^1 and X^2 are the pre-given subsets), has been set to

$$s_{xx'} = (1-\alpha)\,\delta_{j(x)j(x')} + \alpha\, r_{xx'}, \qquad (4.18)$$

where $\delta_{j(x)j(x')}$ – the basic similarity level – is 1 if $x \in X^1$ and $x' \in X^2$ are, by construction, in the same (j-th) coupled cluster or otherwise 0 and $r_{xx'}$ – the random component – is sampled uniformly between 0 and 1, differently for each x and x' in each experiment. The randomness proportion parameter α (i.e. level of added noise), also between 0 to 1, is fixed throughout each experiment, to maintain a steady average noise level.

In order to study the effect of the coupled-cluster inner proportion, we have run four sets of experiments. Given subsets X^1 and X^2 consisting of 32 elements each, four types of synthetic coupled-clustering configurations have been constructed, in which the sizes n_j^1 and n_j^2 of the sub-cluster pairs $c_j^1 \subset X^1$ and $c_j^2 \subset X^2$, together forming the j-th coupled-cluster, have been set as follows:

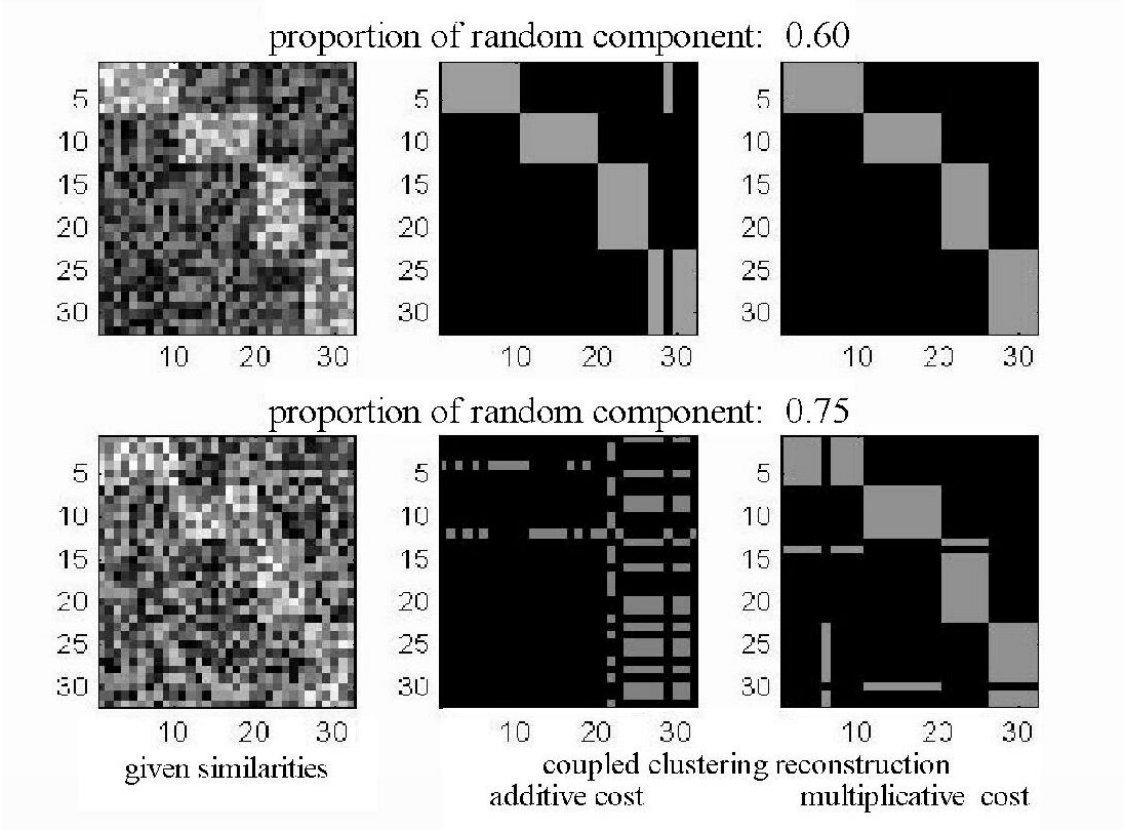

Figure 4.2: Reconstruction of synthetic coupled-clustering configurations of the '10-6 coupling' target configuration type from noisy similarity data. Lines and columns of the plotted gray-level matrices correspond to members of the two sets. On the left-hand side – original similarity values – the gray-level of each pixel represents the corresponding similarity value between 0 (black) and 1 (white). In the reconstructed data, gray level corresponds to average similarity within each reconstructed cluster. The bottom part demonstrates that the multiplicative cost function, H^3, reconstructs better under intensified noise.

(i) $n_j^1 = n_j^2 = 8$, for $j = 1 \ldots 4$;

(ii) $n_j^1 = 10,\ n_j^2 = 6$ for $j = 1,2$ and $n_j^1 = 6,\ n_j^2 = 10$ for $j = 3,4$;

(iii) $n_j^1 = 12,\ n_j^2 = 4$ for $j = 1,2$ and $n_j^1 = 4,\ n_j^2 = 12$ for $j = 3,4$;

(iv) $n_j^1 = 14,\ n_j^2 = 2$ for $j = 1,2$ and $n_j^1 = 2,\ n_j^2 = 14$ for $j = 3,4$.

These four configuration types, respectively labeled '8-8 coupling', '10-6 coupling', '12-4 coupling' and '14-2 coupling', have been used in the four experiment sets.

It is convenient to visualize a collection of similarity values as a gray-level matrix, where rows and columns correspond to individual elements of the two clustered subsets and each pixel represents the similarity level of the corresponding elements. The diagrams on the left-hand side in Figure 4.2 show

two collections of similarity values generated with two different noise levels. White pixels represent the maximal similarity level in use, 1; black pixels represent the minimal similarity level, 0; the intermediate gray levels represent similarities in between. The middle and right-hand-side columns of Figure 4.2 displays clustering configurations as reconstructed by our algorithm, using the additive H^2 and multiplicative H^3 cost functions respectively, given the input similarity values displayed on the left-hand side. Examples from the 10-6 coupling experiment set, with two levels of noise, are displayed. Bright pixels indicate that the corresponding elements are in the same reconstructed cluster. It demonstrates that, for the 10-6 coupling, the multiplicative variant H^3 tends to tolerate noise better than the additive variant H^2 and that this advantage grows when the noise level intensifies (bottom of Figure 4.2).

The performance over all experiments in each set has been measured through the *PURITY* measure, introduced in Section 4.2. Figure 4.3 displays average accuracy for the changing noise levels, separately for each experiment set. The multiplicative cost function H^3, is biased toward balanced coupled clusters, i.e. clusters in which the inner proportion is close to the global proportion of the subsets (which are equal in size in our case). Our experiments indeed verify that H^3 reconstruct better than the other functions, particularly in cases of almost balanced inner proportions.

Figure 4.3 shows that the accuracy obtained using the restricted standard-clustering function H^1 is consistently worse than the accuracy of H^3. In addition, for all internal proportions, there is some range, on the left-hand side of each curve, in which H^3 performs better than the additive function H^2. The range where H^3 is superior to H^2 is almost unnoticeable for the sharply imbalanced internal proportion (2-14 coupling) but becomes prominent as the internal proportion approaches balance. Consequently, it makes sense to use the additive function H^2 only if both: (i) there is a good reason to assume that the data contains mostly imbalanced coupled clusters and (ii) there is a reason to assume high level of noise. Real world data might be noisy, but given no explicit indication that the emerging configurations are inherently imbalanced, the multiplicative function H^3 is preferable. Consequently, we have used H^3 in our experiments with textual data, described in the following sections.

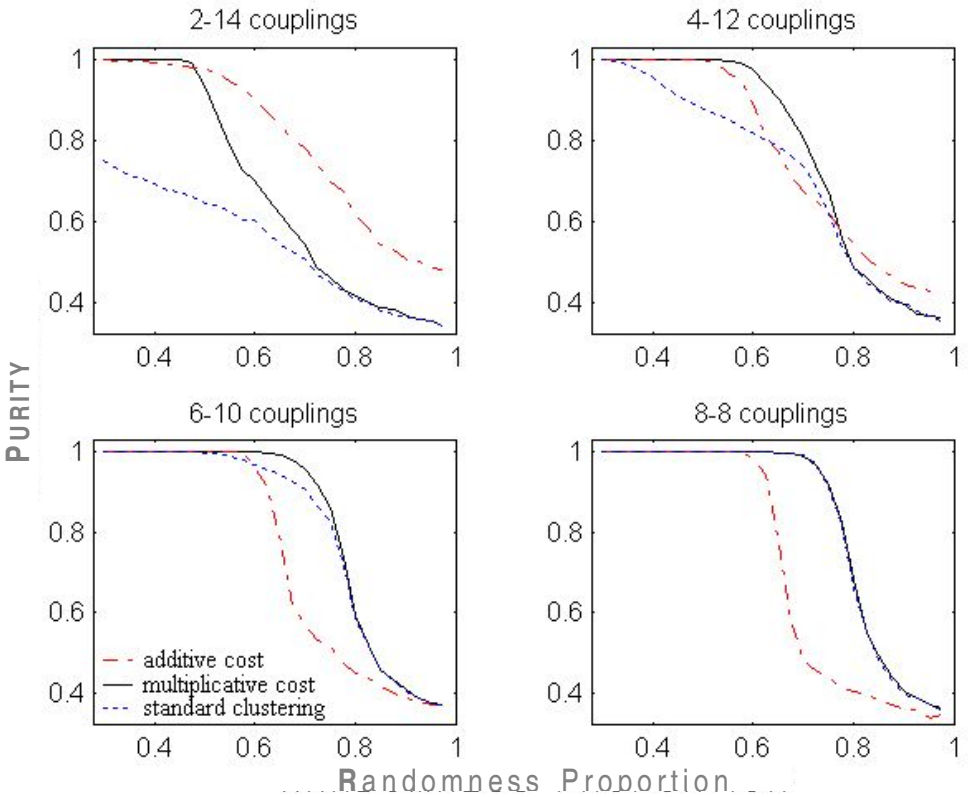

Figure 4.3: Purity as a function of the noise level (randomness proportion) for different coupled size proportions, obtained through experiments in reconstructing synthetic coupled-clustering configurations. For each proportion, results obtained using the straightforward adaptation of the original method (H^1, termed here "standard clustering"), "additive" (H^2) and "multiplicative" (H^3) cost functions are compared.

4.4 Identifying Corresponding Topics in Textual Corpora

In this section, we demonstrate the capabilities of the coupled clustering algorithm with respect to real-world textual data, namely pairs of sets of keywords (the subsets X^1, X^2 mentioned in Subsection 4.2.2) along with counts of co-occurring content words, taking the role of features. The keywords have been extracted from given corpora focused on distinct domains. Our experiments have been motivated by the target of identifying, by means of the induced coupled clusters, concepts that play similar or analogous roles in the examined domains. In Subsection 4.4.1, the keyword sets are extracted from collections of news articles referring to two conflicts of different character that are nowadays in the focus of public attention: the *Middle East conflict* and the dispute over electronic media copyright, demonstrated through the *Napster case*. Our experiments revealed some illuminating correspondences between the two seemingly unrelated conflicts. In Subsection 4.4.2, we turn to larger corpora focused on various religions, specifically *Buddhism*, *Christianity*, *Hinduism*,

Islam and *Judaism*. Hence, the task is to explicate common, or equivalent, aspects of the examined religions. This inter-religion comparison is further analyzed and evaluated also in subsequent sections.

In both conflict and religion comparisons, our setting assumes that the datasets are given or that they can be extracted automatically. We have used the TextAnalyst software by MicroSystems Ltd.[3], which is capable of identifying key-phrases in given text, to generate datasets for our experiments. From the terms and phrases that have been identified by the software, we have excluded the items that have appeared in fewer than three documents. Thus, relatively rare terms and phrases that the software has inappropriately segmented have been filtered out.

After extracting the datasets, between-subset similarities, if not pre-given, are calculated. In general terms, every extracted keyword is represented by a co-occurrence vector, whose entries essentially correspond to the co-located words (concrete examples follow in the subsections bellow), excluding a limited list of function words. Then, between-subset similarity values are calculated using methods, such as those described in Subsection 4.1.2, to adapt the data for the similarity-based coupled-clustering algorithmic setting introduced in Section 4.2. We differentiate between two optional sources that can provide the co-occurrence data for the similarity calculations. One option is to base the calculations on co-occurrences within the same corpora from which the keyword sets have been extracted. Thus, the calculated similarity values naturally reflect the context in which the comparison is being made. This approach has underlay most of the coupled clustering experiments that we have conducted (Subsection 4.4.2). However, sometimes the compared corpora might be of small size and there is a need to rely on a more informative statistical source. An alternative option is to utilize the co-occurrences within an additional independent corpus for the required similarity calculations. In order to produce reliable and accurate similarity values, such independent corpus can be chosen to be significantly larger than the compared ones, but it is important that it addresses well the topics that are being compared, so the context reflected by the similarities is still relevant. This approach, making use of pre-given similarity values, is demonstrated in the following subsection.

4.4.1 Conflict Keyword Clustering Based on Pre-given Similarities

The conflict corpora are composed of about 30 news articles each (200–500 word tokens in every article), regarding the two above-mentioned conflicts: the Middle East conflict and the dispute over music copyright. The articles were downloaded in October 2000.

---

[3] An evaluation copy of TextAnalyst 2.3 is available for download at **http://www.megaputer.com/php/eval.php3**.

Table 4.1: Coupled clustering of conflict related keywords. Every row in the table contains the keywords of one coupled cluster. Cluster titles and titles of the three groups of clusters were added by the author.

	Middle-East	Music Copyright
Parties and 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 and Resources in Dispute		
Locations	**home house street**	**block site**
Protection	**housing security**	**copyright service**
Activity and Procedure		
Resolution	**defeat election mandate meeting**	**decision**
Activities1	**assistance settlement**	**innovation program swap**
Activities2	**disarm extradite extradition face**	**use**
Confrontation	**attack**	**digital infringement label shut violation**
Communication	**declare meet**	**listen violate**
Poorly-clustered keywords		
low similarity values	**interview peace weapon**	**existing found infringe listening medium music song stream worldwide**
no similarity values	**armed diplomatic**	

We have obtained the similarities from a large body of word similarity values that have been calculated by Dekang Lin, independently of our project (Lin, 1998). Lin has applied the sim_L similarity measure (Subsection 4.2, Equation 4.9) to word co-occurrence statistics within syntactic relations, extracted from a very large news-article corpus.[4] We assume that this corpus includes sufficient representation of the conflict keyword sets in relevant contexts. That is: even if the articles in the corpus do not explicitly discuss the concrete conflicts, it is likely that they address similar issues, which are rather typical as news topics. In particular, occurrences of the clustered keywords within this corpus are assumed to denote meanings resembling their sense within our small article collection.

As Table 4.1 shows, the coupled-clusters that have been obtained by our algorithm fall, according to our classification, within three main categories: "Parties and Administration", "Issues and Resources

[4] This corpus contains 64 million word tokens from Wall Street Journal, San Jose Mercury, and AP Newswire. The similarity data is available at **http://armena.cs.ualberta.ca/lindek/downloads/sims.lsp.gz**.

in Dispute" and "Activities and Procedure". To improve readability, we have also added an individual title to each cluster.

The keywords labeled "poorly-clustered", at the bottom of Table 4.1, are assigned to a cluster with average similarity considerably lower than the other clusters, or for which no relevant between-subset similarities are found in Lin's similarity database. Consequently, these keywords could be straightforwardly filtered out. However, poorly clustered elements persistently occur in most of our experiments and we include them here for the sake of conveying the whole picture.

Making use of pre-given similarity data is, on the one hand, trivially advantageous. Apart from saving programming and computing resources, such similarity data typically relies on rich statistics and its quality is independently verified. Moreover: in principle, pre-given similarity data could be utilized for further experiments in clustering additional datasets that are adequately represented in the similarity database. However, there are several disadvantages in taking this route. First, reliable relevant similarity data is not always available. In addition, the context of comparing two particular domains might not be fully articulated within generic similarity data that has been extracted in a much broader context. For example, the interesting case where the same keyword appears in both clustered sets, but it is used for different meanings, could not be traced. A keyword used differently in distinct corpora would co-occur with different features in each corpus. In contrast, when similarities are computed from a unified corpus, self-similarity is generally equal to the highest possible value (1 in Lin's measure), which is typically much higher than other similarity values. In such case, the two distinct instances of a keyword presenting in both clustered sets would always fall within the same coupled-cluster.

4.4.2 Religion Keyword Clustering

This subsection introduces the main body of our data, to which, from this point on, the coupled clustering method is systematically applied, followed by detailed examination and evaluation of the outcome. The same data is further analyzed, in the next chapter, through additional algorithmic extensions. The data consists of five corpora, each focusing on a different religion: Buddhism, Christianity, Hinduism, Islam and Judaism, to which we apply our methods in order to compare the religions to one another and to identify corresponding aspects. The corpora used here significantly extend the ones used by Marx, et al. (2002).

As we have noted earlier, one of the options for inducing input similarity values is by using co-occurrence statistics from corpora that are focused on the compared domains. These can be the same corpora from which the clustered keywords are extracted. In such case, it is clear that each keyword appears in its relevant sense or senses. Hence, context dependent subtleties, such as identical

keywords denoting different meanings, can be revealed. In this case, we rely on the assumption that there is a substantial overlap between the features, namely words commonly co-occurring in the two corpora, and that at least some of the overlapping features are used similarly within both. Specifically, we assume that the corpora to which we refer below – introductory web pages and encyclopedic entries concerning religions – contain enough common vocabulary directed towards some "average-level" reader, thus enabling co-occurrence-based similarity calculations that are fairly informative. In summary, while pre-given similarity data might typically result from richer statistics over a unified set of features, the alternative might fit better the context of the task at hand, but depends on rich enough statistics of shared features.

4.4.2.1 The Data

The religion data consists of five corpora containing encyclopedic entries, electronic periodicals and additional introductory web pages that were downloaded from the Internet. The five corpora contains 1.5–2 million word tokens (8.5–13 Megabyte) each. Using the TreeTagger[5] software, we have filtered out all function words according to their part-of-speech (POS) and substituted each one of the remaining words by its lemma. This way, each corpus has been shrunk to 0.8–1.2 million tokens (5.5–8.5 Megabyte). More details about the corpora can be found in Appendix A. In addition to the keywords extracted by TextAnalyst software (described above), the elements of the clustered sets include keywords that have been provided by comparative religion experts (the data provided by experts has been primarily used for quantitative evaluation, see Subsection 4.4.2.3). The total size of each of the final keyword sets is 180–240, of which 15–20% were not extracted by TextAnalyst but solely by the experts. Each keyword is represented by its co-occurrence vector, as extracted from its own corpus (so the same keyword that is relevant to two or more corpora has different representation with respect to each corpus). In counting co-occurrences, we have used two-sided sliding window of ±5 words, truncated by sentence ends (similarly to Smadja, 1993). On one hand, this window size captures most syntactic relations (Martin, Al & van Sterkenburg, 1983). On the other hand, this scope is wide enough to score terms that refer to the same topic in general – and not only literally interchangeable terms – as similar (Gorodetsky, 2001), which accords our aim of identifying corresponding topics. Appendix A contains details of some of the features that are most common in the corpora. The keyword sets are introduced through some of their items along with exemplifying features and corresponding co-occurrence counts.

Each one of the clustered keywords is represented by a (sparse) vector, whose entries are the counts of the keyword's co-occurrences with each feature. We have applied to the obtained vectorial

[5] TreeTagger – a language independent POS tagger and lemmatizer – is available for download from http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/DecisionTreeTagger.html.

representations the sim_{DMM} similarity measure, which incorporates detailed information on the data (Dagan, Marcus & Markovitch 1995; Subsection 4.1.2, Eq. 4.8; sim_L is less detailed in that it only distinguishes between features that are present in one of the two vectors: features that are present in both do not contribute to the dissimilarity even though the values might be different, while sim_{DMM} utilizes value differences for features that are positive in both vectors). After calculating between-subset similarities, we ran the coupled clustering algorithm on each pair of subsets.

4.4.2.2 Qualitative Overview of the Results

Appendix B contains a detailed sample of coupled-clustering results, with number of clusters $k = 16$, including four religion pairs: Buddhism–Christianity, Christianity–Hinduism, Hinduism–Islam and Islam–Judaism. The coupled keyword clusters are ordered by their average similarity in descending order. The poorly clustered elements – those in the 16th cluster with the lowest average similarity – are not shown. We have attached intuitive titles to each cluster for readability and orientation.

The following post-hoc thematic analysis combines careful examination of the results described in this section and some background reading (particularly Smart, 1989; see discussion in Section 4.5). The obtained clusters (and additional results that are not shown concerning other religion pairs) appear to reflect consistently several themes that share commonalities across the different subsets:

∗ The religious experience:

  This theme incorporates, for instance, terms referring to spiritual and mental conditions that lead to, or are the result of, the search for the religious message or religious belief.

∗ Theology and philosophy:

  This covers several aspects such as ethics and other basic principles to be followed. Two notable subtopics that are typically expressed through distinct clusters: qualities and attributes that are admired, usually related to the nature of the divine and, in contrast, the issues of sorrow, suffering, sin and punishment, usually adjoined with terms referring to the reward promised to those who do not violate the religious way.

∗ Institutional organizations:

  This topic covers the various schools and traditions within the religion and, sometimes, the history of their development. It includes, for instance, terms referring to various types of priests serving at the particular religion. It often involves names of places, where the various traditions have been originated, or are currently practiced.

∗ Practice and custom:

  This topic includes ritual aspects of the religion, for instance, terms referring to dietary rules, pilgrimage, holy places and festivals.

∗ The scriptures:

In addition to the names of holy writings, there are two notable sub-themes: terms referring to teaching and scholarship and terms related to myths and narratives. The latter sub-theme often involves names of figures and places that are related with the narratives.

Furthermore, keyword coupled clusters are a valuable source for additional specific analyses of varied sorts. We shall exemplify this here only briefly:

- **Religion founders.** The names of the central figures of each religion are often clustered together within the same coupled-cluster. Whenever such a cluster is not focused on other personal names, the cluster's terms often convey prominent attributes of the central figure, thus provide in some sense the key to the whole religion – the ideal attribute it adopts. Examples are *being*, *practice* and *teaching* with regard to Buddha, *believing*, *faith*, *love* regarding Jesus Christ and *messenger*, *prophet*, *Quran* regarding Mohamed.
- **Family relations.** Family related terms – *family*, *wife*, *husband*, *marriage*, *mother*, *sister*, *brother*, *daughter*, *child* (excluding *father* and *son*, which often convey additional meaning) – are distributed differently over clusters, in different pairs of religions. Islam plays a pivotal role with relation to these terms. Clustering Islam terms against those of Christianity and Judaism, the family terms are concentrated within a single coupled cluster. This provides a hint regarding the central part of family issues in the Islam, when it is viewed in light of the other western religions where comparable aspects are present (in contrast, comparing Christianity and Judaism with one another, the same terms are distributed among four different couple clusters). When Islam is compared to Buddhism, the family-related terms are divided among varied contexts: "personal relationships" (*enemy*, *fight*, *meet*, *responsibility*, …), "sin and prohibitions" (*forbid*, *kill*, *pain*, *punishment*, …), "figures in narratives" (*Abraham*, *Ishmael*, *Moses*, *caliph*, *tribe*, …).

We provide further qualitative evaluation of more results concerning comparison of religions in the following subsections.

4.4.2.3 Expert Data Used for Evaluation

As the previous section explicates, our empirical experiments have been concentrated on the comparative study of religions. In fact, the existence of such a discipline and its presumed potential as a source for external standards was an initial reason for applying our framework for religion comparison. The different religions are non-trivially related to one another. Thus, religions seem to be liable to varied types of analysis and views that might underlie interesting correspondences between them. Our working hypothesis is that corresponding aspects of distinct religions would be expressed through common features, i.e. commonly shared content words, implying close, or related,

meaning. However, from the data contributed by the experts it also becomes apparent that there is no precise definition or consensual agreement with regard to the aspects of correspondence and the particular terms capturing them.

We have asked human subjects, whose academic field of expertise is the comparative study of religions, to perform manually the coupled clustering task in the following manner. The experts have been asked to explicate the most prominent equivalent aspects common to given pairs of religions. (To convey a broad notion of equivalency, we have included the following phrase in the instructions: "… features and aspects that are *similar*, or *resembling*, or *parallel*, or *equivalent*, or *analogous* in the two religions under examination…"; see Appendix C for the full instructions). Then, the experts have been requested to specify representative terms that characteristically address each one of the identified similar aspects, within the content world of the two compared religions. The resulting pairs of corresponding sets, or *classes*, of terms, each of which addressing one aspect of similarity between two compared religions, form our external standard – a *class configuration* (see Section 3.3). Such a configuration is used for evaluating our results regarding the particular religion pair to which it refers.

The task of explicating keywords, dissociated from any wider context, was new and somewhat unusual to our experts. We have made efforts not to add on top of that any bias that further restrictions might cause. We have provided only rough guidelines regarding the number or content of equivalent aspects (i.e. expert classes), and the number and identity of terms that are associated with each aspect within each religion (i.e. the size of the coupled clusters; see Appendix C). We have not set limits to the number of equivalent aspects with which any word can be associated.

We have got responses from four people that have accomplished the task – two graduate students and two university professors, from Finland, Israel and New Zealand. Perhaps due to the pretty permissive guidelines, we could not use some parts of the contributed data. There were several terms that did not occur, or occurred rarely (i.e. less than 40 times in the relevant corpus), in our corpora. A particular contribution, by one of the four experts, contained too few prevalent terms and thus has been discarded altogether, so we have been left with the expert class configurations contributed by the three remaining experts. One of the experts has provided comparisons between all 10 possible pairs of the five religions. Another expert has provided the following comparisons: Buddhism-Christianity Buddhism-Hinduism, Christianity-Islam and Christianity-Hinduism. The third expert has provided the three possible comparisons among Christianity, Islam and Judaism.

The data in use still included phrases conveying ideas that were far too composite than what we expected from key terms (e.g., the phrase *not-admiring-something-that-belongs-to-someone-else*). Most phrases of this type were excluded. With regard to few of them we made some editorial

interpretations. For example (one of the most extreme cases): identifying the single expert phrase *obeying-God-and-not-mentioning-his-name-without-a-reason* with co-locations of the terms *obey* and *God*. Extensive work has been required also for other editing tasks, such as identifying alternating spellings. In total, we discarded from the expert classes about 50 (10%) of the terms contributed by the three experts, so that a total of 448 terms were left to be used in the evaluation. The detailed expert data is described further in Appendix C.

As noted, the field of comparative study of religions does not lend itself to an absolute measure for assessing keyword-grouping results, through the experts' contribution. Accordingly, our external standard cannot be regarded as conveying a definite academic directive. Nevertheless, we suppose that our appliance to domain experts have yielded data that is both more reliable and richer than what could have been collected from, say, "educated human subjects". Note that even for experts the task of identifying corresponding themes in religions is intrinsically subjective, most likely because their skills do not underlie clear-cut criteria. The fact that we have used data from three participants allows us to provide some notion of the maximal level of precision that we can, in principle, get.

4.4.2.4 Examples of Expert Data versus Coupled Clustering Output

Table 4.2 shows concrete examples for our results in comparison to the expert data used for evaluation. The top of the table (A) displays a specific coupled class configuration contributed by one of the experts, pertaining to Christianity to Hinduism. The expert configuration is followed by the output of our method. The next two configurations, in (B) and (C), have been produced by our coupled clustering algorithm, making use of the multiplicative H^3 and additive H^2 cost functions respectively. Although the expert configuration consists of five clusters, the most convincingly interpretable results, shown in the table have been obtained with eight clusters. The table demonstrates that reconstruction of the expert configuration follows, in several cases, the right direction, but it is still imperfect. There are several expert classes – e.g., the one titled "mysticism" – for which no trace is found in the various computerized outputs. On the other hand, computerized configurations display some level of topical coherence, unrelated with the expert clusters, for example, the cluster that we have titled *religious experience* in Table 4.2 (B), referring to the multiplicative function performance. The "one-to-many" coupled-clusters produced by the additive cost function (C) do reveal, as well, some interesting themes: symbols, doctrine, theological principles and so on. The themes captured, however, are not balanced well over the two religions and consequently do not overlap well with the expert classes, even in cases where there is some thematic correspondence. For comparison with standard single-set clustering, we used the IB method reviewed in the next chapter (Section 5.2). The standard clustering results in Table 4.2 (D), place all Hinduism terms in one cluster, which disallow the detection of any correspondence between the two religions.

Table 4.2: Examples of the expert data and coupled clustering results.

(A) Expert class configuration: The class configuration, comparing Christianity to Hinduism, contributed by expert I. The titles are those given by the expert.

Christianity	Hinduism
1. Scriptures	
new_testament old_testament apostle bible john luke matthew paul revelation	Gita mahabharata upanishad vedas
2. Beliefs and Ideas	
jesus_christ love_of_god devil god cross fish heaven hell resurrection trinity	holy_people trimurti moksha atman brahman reincarnation
3. Society and Politics	
catholic church minister monk priest protestant rome vatican	brahmin caste sadhu
4. Establishments	
bishop cardinal church pope priest	caste gift priest temple
5. Mysticism	
eucharist crucifixion love miracle saint suffer	ashram chakra darshan guru yoga

(B) Multiplicative cost function: Eight-cluster configuration produced by our program with the multiplicative function (the best score was obtained for this eight-cluster configuration, although the expert specified five classes). The eighth cluster, of lowest average intra-cluster similarity, is omitted from the multiplicative cost function results. The remaining seven clusters are shown in full. The results are shown in full, including terms not used by the expert. (Cluster titles are by the author).

Christianity	Hinduism
(1. religious experience)	
being believing child death earth faith father find **god**[2] hear holy jesus **love**[5] man people prayer problem sin soul spirit **suffer**[5] word	being child death family find god human man people soul
(2. writings–1)	
america **bible**[1] book **church**[3,4] evangelical history religious **rome**[3] theology tradition write	ancient art author book christian country history india language philosophy question religion religious sacred sanskrit school science shri south study **temple**[4] tradition **vedas**[1] west write
(theology)	
divinity doctrinal experience human moral relationship religion spiritual	animal attain **brahman**[2] consciousness dharma discipline divinity existence experience faith freedom idea karma law liberation practice ritual sense shiva social society spirit spirituality teach universe word **yoga**[5]
(writings–2)	
author chapter greek hebrew **luke**[1] **matthew**[1] **new_testament**[1] **old_testament**[1] passage **revelation**[1] scripture study text theory translate writer writing	epic **gita**[1] hymn literature **mahabharata**[1] purana ramayana rigveda scripture story sutra teaching text **upanishad**[1] writing
(doctrine / schools)	
ancient baptist **bishop**[4] **catholic**[3] constantinople convert council establish found german jew luther organization orthodox **pope**[4] **protestant**[3] university **vatican**[3] west	aryan authority **brahmin**[3] buddhism **caste**[3,4] civilization doctrine found founder jain muslim scholar shaiva
(tradition / cutoms)	
christmas city disciple family friend home house jerusalem learn meet member **minister**[3] ministry school service sunday woman worship	**ashram**[5] ceremony dance festival ganesh **gift**[4] holy krishna learn meditation pilgrimage prayer **priest**[4] puja rama **sadhu**[3] son star student teacher
(narratives)	
abraham angel **apostle**[1] authority baptism baptize believer birth bless blood command confess **devil**[2] eat eye face faithful fire flesh forgiveness gift gospel grant **heaven**[2] **hell**[2] holy_spirit israel **jesus_christ**[2] **john**[1] judgment kingdom law listen mankind moses mother **paul**[1] pay peace preach prophet punishment question redemption refer repentance **resurrection**[2] reward righteousness sabbath sacrifice **saint**[5] sake salvation savior sinful sinner teach voice water win	birth devotee earth **guru**[5] heaven mother person sacrifice

Table 4.2 (cont.): Examples of the expert data and coupled clustering results

(C) Additive cost function: Eight-cluster configuration produced by our program with the additive function. All clusters are displayd, but terms not used by the expert are shown only in the cases they are the only ones in their cell. Expert terms are in bold font. Superscripts inticate expert class number. (Cluster titles are by the author).

Christianity	Hinduism
(1. spirituality)	
spiritual	**guru**[5] **yoga**[5]
(2. religious)	
religious	**ashram**[5] **brahmin**[3] **caste**[3,4] **priest**[3,4] **temple**[4]
(3. personal experience)	
apostle[1] **bible**[1] **devil**[2] **god**[2] **hell**[2] **jesus_christ**[2] **john**[1] **love**[5] **love_of_god**[2] **paul**[1] **resurrection**[2] **suffer**[5]	Person
(4. history – writings)	
history	**gita**[1] **mahabharata**[1] **upanishad**[1] **vedas**[1]
(5. establishments)	
church[4] **minister**[3] **saint**[5]	Devotee
(6. symbols)	
cross[2] **fish**[2] **heaven**[2] **miracle**[5]	Heaven
(7. doctrine)	
bishop[4] **cardinal**[4] **catholic**[3] **crucifixion**[5] **eucharist**[5] **luke**[1] **matthew**[1] **monk**[3] **new_testament**[1] **old_testament**[1] **pope**[4] **priest**[3,4] **protestant**[3] **revelation**[1] **rome**[3] **vatican**[3]	doctrine
(8. theological principles)	
trinity[2]	**atman**[2] **brahman**[2] **chakra**[5] **darshan**[5] **gift**[4] **holy_people**[2] **moksha**[2] **reincarnation**[2] **sadhu**[3] **trimurti**[2]

(D) Single set clustering: Eight-cluster configuration produced by a standard clustering method (the information bottleneck iterative algorithm, producing soft clusters (Section 5.2); each term is assigned into its most probable cluster). The Hinduism terms were all assigned in one cluster, so only Christianinity terms are shown. Exemplifying terms not used by the experts are shown only in the cases where there have been no expert terms in their cluster. Expert terms are in bold font. Superscripts indicate expert class number. (Cluster titles are by the author).

Christianity	Hinduism
(1. establishment-A)	
vatican[3] **pope**[4] **cardinal**[4] **rome**[3] **bishop**[4] **catholic**[3] **protestant**[3]	
(2. customs/<general for Hinduism>)	
trade pilgrimage	**[All Hinduism terms were assigned to cluster 2]**
(3. spirituality)	
holy_spirit holy reign spirit	
(4. establishment-B)	
church[4] **monk**[3] **eucharist**[5]	
(5. writings/figures)	
luke[1] **matthew**[1]	
(6. doctrine)	
postmodern theology luther evangelical ethic tradition religious religion founder	
(7. <general>)	
apostle[1] **bible**[1] **devil**[2] **god**[2] **hell**[2] **jesus_christ**[2] **john**[1] **love**[5] **love_of_god**[2] **paul**[1] **resurrection**[2] **suffer**[5] **minister**[3] **saint**[5] **cross**[2] **fish**[2] **heaven**[2] **miracle**[5] **crucifixion**[5] **priest**[3,4] **revelation**[1] **trinity**[2]	
(8. sacred writings)	
new_testament[1] **old_testament**[1]	

4.4.2.5 Quantification of the Overlap with the Expert Data

We employed the Jaccard coefficient to quantify the overlap between our output clusters and the classes provided by the experts. We used the version of Jaccard coefficient that is specifically adapted to the coupled clustering task, considering only cross-dataset element pairs (*JACCARD-PROB-CP*; see Chapter 3, Subsection 3.3.2.2). In order to lay common grounds for measuring the overlap, we eliminated from our clusters those terms that were not used by the expert. Note that the data was clustered in full, so that the terms not used by the expert were deleted from the outcome, *after* the clustering process was completed. This procedure differs from the one used by Marx et al. (2002), where the clustering algorithm was applied to partial datasets that included only expert's terms. Considering the noisy impact of those many items that are not in the target classes (about 400 items per couple of full datasets, compared to an average of 54 expert items used in each evaluation), the current procedure demonstrates a higher degree of robustness.

We compared coupled clustering results obtained with the multiplicative cost function H^3 to the additive function H^2, as well as to random assignments and to the clusters produced by standard clustering method – the *information bottleneck* iterative algorithm (Tishby, Pereira & Bialek, 1999; reviewed in Chapter 5, Section 5.2) – applied to the union of the two coupled subsets. The information bottleneck method produces soft probabilistic clustering, i.e., it assigns each element to all clusters with probabilistic assignment values that sum up to 1. We turned these probabilistic assignments into hard ones, by considering each element as if it is assigned into its most probable cluster. The original soft IB clusters can be evaluated through the methods we use as well, but in general, they score somewhat worse than the hard version.

Figure 4.4 displays the results for all 17 evaluation cases examined. High Jaccard coefficient values imply high degree of overlap with the expert classes. The number of clusters indicated by each expert, which is denoted in the figure by a dotted vertical line, does not perfectly predict the number of clusters that actually obtain the highest score, so it cannot be assumed known in advance. We rather examine output configurations with numbers of clusters that vary over a reasonable predetermined range: two to 16 clusters. Averaged over all shown cluster numbers across all religion pair cases, the differences between the methods are all statistically significant, except for the difference between the additive cost function and the standard single-set clustering. Particularly, Figure 4.4 exhibits the superiority of the multiplicative cost in the vast majority of cases, over the whole cluster number range. In some cases, additionally to the highest scoring configuration, there are local picks (maxima) along the result graph, indicating that there is more than one meaningful interpretation to the data, corresponding to different levels of detail.

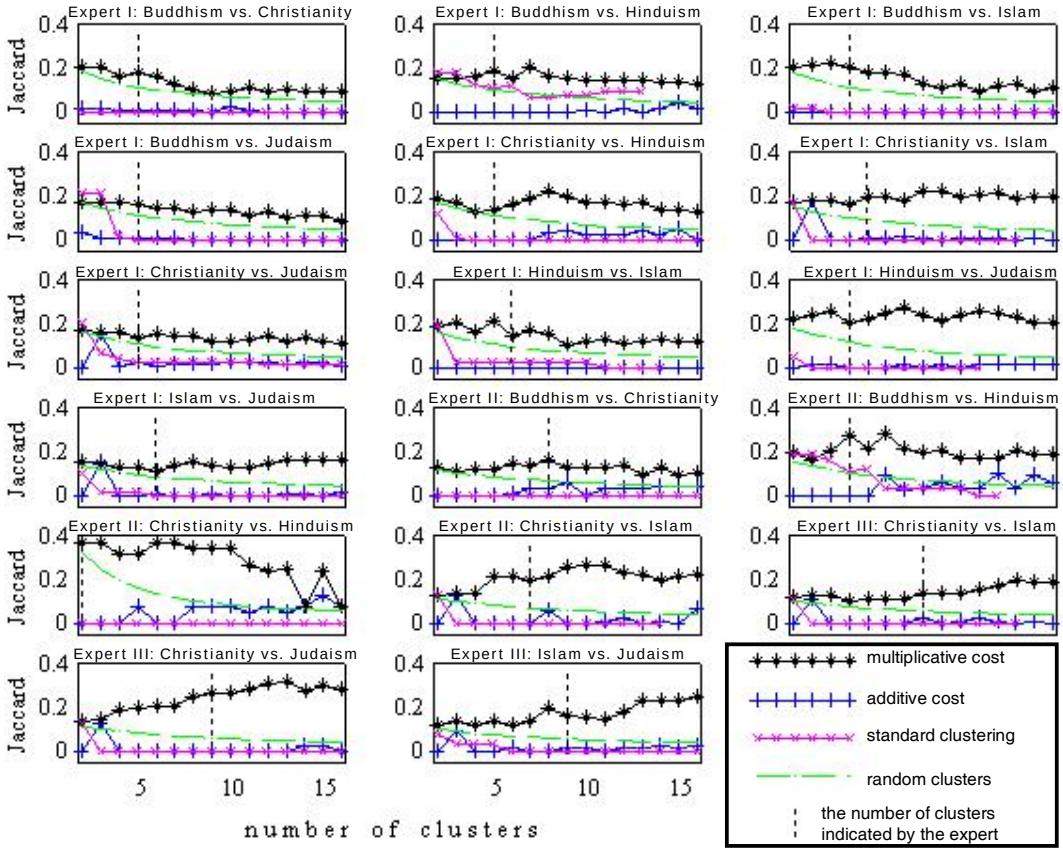

Figure 4.4: The religion keyword coupled-clustering results evaluated relatively to the expert classes. Jaccard scores are shown for cluster numbers range from two to 16, for all 17 cases: ten by expert I, four by expert II and three by expert III. The methods in use: coupled clustering with the multiplicative and additive cost functions, and a standard clustering method (information bottleneck). Scores of random assignments are shown as well.

In most cases, the additive-cost and standard-clustering Jaccard scores shown in Figure 4.4 lie below the corresponding random-assignment scores. The reason is that the version of Jaccard coefficient used here considers only cross-dataset element pairs (Chapter 3, Subsection 3.3.2.2). The additive cost function tends to form "one-to-many" coupled clusters, i.e., clusters that contain only one, or very few, elements from one of the datasets (for an example, see Table 4.2 (C)). Standard clustering, as well, tends to follow within-dataset regularities, inducing similarly imbalanced clusters.

4.4.2.6 Agreement between the Experts

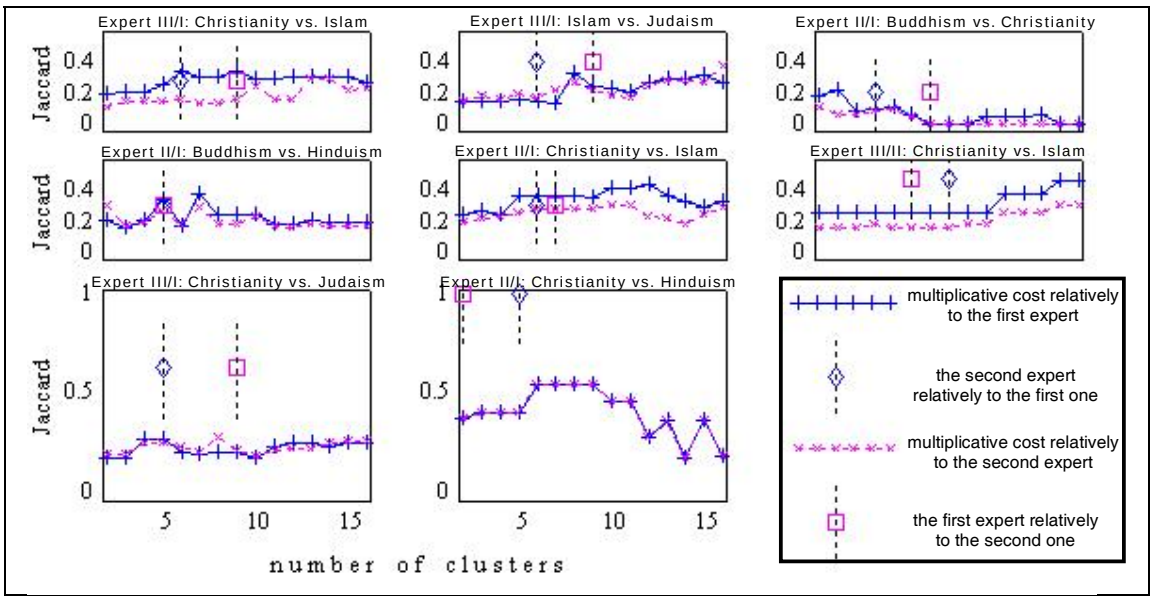

Figure 4.5: The religion keyword coupled-clustering results evaluated on partial sets of terms: those used by two experts for the same cross religion comparison. The Jaccard scores of the 10 cases are shown for all cluster numbers from two to 16, three common to experts I and II, three common to experts I and III and one common to experts II and III. For comparison, we show the level of agreement between the experts, i.e., the Jaccard score each expert configuration achieves in approximating the classes provided by the other expert.

In this subsection, we quantify the subjectivity level that can be ascribed to the evaluation criterion in use and we examine the portion of our results comparable with the limits set by this subjectivity level. There was one religion pair (Christianity/Islam) to which all three experts generated evaluation data independently and five additional religion pairs to which two experts generated data independently (Buddhism/Christianity, Buddhism/Hinduism and Christianity/Hinduism by expert I and II; Christianity/Judaism and Islam/Judaism by experts I and III). Together, evaluating data of an expert against data regarding the same religion pair contributed by another expert gave a total of 16 evaluation cross-expert evaluation cases: ten cases resulting from the religion pairs addressed by two experts and six cases from the pair addressed by all experts, as this religion pair was actually addressed by three pairs of experts. Note that each expert pair was judged twice, taking one expert as a gold standard and the other expert as the one being evaluated.

Evaluating expert data through comparing it to data of another expert measures cross-expert overlap over the set of terms used in common by both experts, thus such evaluation results provide an indication for the level of agreement between the experts. Figure 4.5 displays the Jaccard scores indicating these cross-expert agreement levels for each pair of religions, along with the results

obtained by our method on the same small sets of commonly used terms. In order to set common grounds to our results with cross-expert agreement, after the clusters were formed the terms not used by *either* expert were discarded from our clusters, so that the term sets left are much smaller than the ones used for evaluation in the previous subsection. The figure shows that our results approach the cross-expert agreement level in some of the cases. However, the averaged results in table 4.4 show that even the best clustering results for each case (over the range of number of clusters) are significantly inferior to the agreement between experts. In the next chapter, these results would be significantly improved.

Table 4.4: Quantitative measures for cross-expert agreement, obtained through applying the evaluation methods to expert-classes using another expert class configuration as a criterion. For comparison, we show mean scores of our results with respect to the 16 corresponding religion comparison cases, to which an additional expert has referred. The evaluation is restricted to the items common to both experts. In parentheses: the average over the best score of each case.

	Jaccard Coefficient
Means ± standard deviations **for cross-expert agreement quantitative assessment**	
Cross-expert Agreement	**0.450**±0.249
Means ± standard deviations **over the 16 cross-expert scores averaged** (best) **over all examined numbers of clusters 2–16**	
Multiplicative Cost	**0.237**±0.101 (0.344±0.114)
Difference: Expert – Multiplicative	**0.214**±0.194 (0.106±0.193)

4.5 Discussion

In this chapter we have formalized and implemented the coupled clustering problem that was introduced in general terms in the previous chapter: clustering two pre-given element subsets to matching parts so that each matched pair forms a coupled cluster. Formalization of the task took place in the familiar pairwise cost-based data clustering framework (Subsection 4.2.2). The implementation has used the stochastic Gibbs Sampler search method (Subsection 4.2.4). The requirement of matching the formed subset parts has been realized through restricting the pairwise clustering setting to only those similarities between members of distinct subsets (Subsection 4.2.1).

The results demonstrate that our approach addresses the coupled clustering task fairly well, not only with respect to tailored synthetic task (Section 4.3), but also for tackling an interesting real-world problem (Section 4.4). Neither standard clustering techniques nor simplistic approach, such as the one suggested by the straightforward additive cost function (Eq. 4.11), address the examined task as

well as the solution dedicatedly designed for the problem, namely the multiplicative cost function (Eq. 4.14).

The expertise of the individuals that participated in creating our evaluation criteria did not completely eliminate the subjectivity inherent to the task of identifying and matching terms related with various religions. Given the inherently subjective task and the lack of clear-cut criteria for matching equivalent terms or themes within the studied data, we find the results encouraging and inherently not too far from the level of agreement among the experts.

Yet, several aspects in the method that we have introduced seem as non-negligible limitations. There is a source of information, the between-subset similarities, which are not utilized at all. Should they really play no role in the formed correspondence between systems aligned with respect to each other? Another point to note is that the conversion from element-feature data to pairwise similarities (through methods as the ones described in Subsection 4.1.2) implies additional loss of information.

In the next chapter of this work, we introduce another method that generalizes the coupled clustering setting across several aspects and, in addition, addresses the points we have mentioned.