Heterogeneous Data Co-Clustering by Pseudo-Semantic Affinity Functions

Alberto Messina and Maurizio Montagnuolo

RAI Centre for Research and Technological Innovation,
Corso E. Giambone 68, I10135 Torino, Italy
{a.messina,maurizio.montagnuolo}@rai.it

Abstract. The convergence between Web technology and multimedia production is enabling the distribution of content through dynamic media platforms such as RSS feeds and hybrid digital television. Heterogeneous data clustering is needed to analyse, manage and access desired information from this variety of information sources. This paper defines a new class of pseudo-semantic affinity functions that allow for a compact representation of cross-modal documents relations.

Keywords: Co-clustering, Pseudo-semantic Model, Syntactic Affinity

1 Introduction and Related Work

The proliferation of multimedia production tools is enabling the convergence of different media technologies. The result of this technological breakthrough is the creation of new relationships between different media objects and the modalities in which they are generated and consumed.

Heterogeneous data clustering can be defined as the set of methods that combine data assets that are different in nature, for example audiovisual and textual material, and use the obtained information to present additional knowledge, potentially not discoverable by the analysis of the individual sources. A type of multi-dimensional heterogeneous data clustering is co-clustering, which allows simultaneous clustering of the rows and columns of a matrix. Given a set of M rows in N columns, a co-clustering algorithm generates co-clusters i.e. a subset of rows which exhibit similar behaviour across a subset of columns, or vice-versa. A pioneer work is described in [1], where a collection of text data is modelled as a bipartite graph between documents and words, thus reducing the clustering problem to a graph partitioning problem. A similar approach applied to multimedia collections is presented in [7]. The work in [5] describes a co-clustering algorithm that monotonically increases the preserved mutual information by intertwining both the row and column clustering at all stages. As clustering of high-dimensional data can suffer from the curse of dimensionality problem, Domeniconi et al. [2] proposed a technique to generate clusters in subspaces spanned by different combinations of dimensions.

Another problem related to co-clustering is how to set the optimum similarity function for maximising the cluster separability [3]. This paper presents an asymmetric vector affinity function on which pseudo-semantic dependency graphs among data objects are built based on the cross-projection of two sets of heterogeneous data spaces. Under this framework we implemented a novel co-clustering algorithm for the seamless fusion of multimedia content coming from various sources. Because of asymmetric nature of the adopted model, hierarchical relations, i.e. syntactic entailment and equivalence, between data objects, as well as representativeness degrees of the found groupings can be expressed.

The paper is organised as follows. §2 describes the mathematical principles on which the pseudo-semantic affinity measure is built. §3 outlines the functional blocks performing the proposed co-clustering algorithm. §4 presents the experimental evaluations conducted to demonstrate the effectiveness and potential of the framework and the algorithms. §5 provides an example of a real-world application based on the proposed technology. §6 concludes the paper.

2 Pseudo-semantic Affinity

In our work co-clustering is based on the on the concept of *pseudo-semantic affinity* (PSA), a generalisation of the concept of *semantic similarity* originally proposed in [4]. By this generalisation we aim at increasing completeness without loosing accuracy. The PSA is loosely inspired by common language when we speak about relationships among people. There we call affine people who are related by some relativeness path. Some affinity concepts are symmetrical, e.g. *being brother of*, some of the symmetrical affinities are similarities, e.g. DNA sequence alignment measurements, but there are also generic affinities in which neither of the two properties hold, e.g. a nephew is not his uncle's uncle.

Definition 1. *Let* $\mathbf{a}, \mathbf{b}$ *be two non-negative real vectors of size N. A pseudo-semantic affinity is a function* $S_{m,n}(\mathbf{a}, \mathbf{b}) : \Re_0^{+N} \times \Re_0^{+N} \Rightarrow \Re_0^{+}$ *of the form:*

$$S_{m,n}(\mathbf{a}, \mathbf{b}) = \frac{\langle \mathbf{a}, \mathbf{b} \rangle}{\|\mathbf{a}\|^m \|\mathbf{b}\|^n} \tag{1}$$

where $m, n \in [0, +\infty)$*, and* $\langle \cdot \rangle$ *is the inner product.*

It can be demonstrated that the PSA is a *pseudoquasimetric*[1] (or hemimetric) for the similarity of $\mathbf{a}$ and $\mathbf{b}$ in the considered vector space. Also, if $\mathbf{a}$ and $\mathbf{b}$ are two feature vectors representing two data objects, then Eq. (1) models syntactic properties of the relations between $\mathbf{a}$ and $\mathbf{b}$, such as equivalence, entailment (or, conversely, dependency), and inclusion, thus providing hierarchical relationships among the considered data sets.

Let be $\Delta = m - n$ and $k \in \Re$, then $S(\cdot)$ has the following analytical properties:

Auto-affinity (2), Co-symmetry (3), Symmetrisation (4), Index Shift/Scaling (5 - 6), Argument Scaling (7), Cosine Generalisation (8)

$$S_{m,n}(\mathbf{a}, \mathbf{a}) = \|\mathbf{a}\|^{1-m} \|\mathbf{a}\|^{1-n} = \|\mathbf{a}\|^{2-n-m} \tag{2}$$

[1] See `http://en.wikipedia.org/wiki/Metric_(mathematics)#Pseudoquasimetrics`

$$S_{m,n}(\mathbf{a},\mathbf{b}) = S_{n,m}(\mathbf{b},\mathbf{a}) \tag{3}$$

$$S_{m,n}(\mathbf{a},\mathbf{b}) = S_{n,n}(\mathbf{a},\mathbf{b})\|\mathbf{a}\|^{-\varDelta} = S_{m,m}(\mathbf{a},\mathbf{b})\|\mathbf{b}\|^{\varDelta} \tag{4}$$

$$S_{m+k,n}(\mathbf{a},\mathbf{b}) = S_{m,n}(\mathbf{a},\mathbf{b})\|\mathbf{a}\|^{-k} = S_{m+k,n+k}(\mathbf{a},\mathbf{b})\|\mathbf{b}\|^{k} \tag{5}$$

$$S_{m,m}(\mathbf{a},\mathbf{b}) = S_{n,n}(\mathbf{a},\mathbf{b})\|\mathbf{a}\|^{-\varDelta}\|\mathbf{b}\|^{-\varDelta} \tag{6}$$

$$S_{m,n}(\mathbf{a},k\mathbf{a}) = k^{1-n}S_{m,n}(\mathbf{a},\mathbf{a}) \tag{7}$$

$$S_{m,n}(\mathbf{a},\mathbf{b}) = \cos(\mathbf{a},\mathbf{b})\|\mathbf{a}\|^{1-n-\varDelta}\|\mathbf{b}\|^{1-n} \tag{8}$$

Eq. (2) implies that the affinity of a vector with itself depends on its norm. However, it is possible to demonstrate that independence from $\|\mathbf{a}\|$ is achieved iff $S_{m,n}(\mathbf{a},\mathbf{a}) = \gamma$, where γ is a constant value. This implies that $\|\mathbf{a}\|^{(2-m-n)} = \gamma$, which is true if and only if $2-m-n = 0$. In this case the PSA is said *well-defined*. Eq. (8) states that any PSA is a generalisation of the Cosine similarity measure, where $\cos(\mathbf{a},\mathbf{b}) = \frac{\langle \mathbf{a},\mathbf{b}\rangle}{\|\mathbf{a}\|\|\mathbf{b}\|} = S_{1,1}(\mathbf{a},\mathbf{b})$.

2.1 Geometrical Properties of PSA

Let $\mathcal{A} = \{\mathbf{x_1},\ldots,\mathbf{x_M}\} \subseteq R_0^{+N}$ be a set of non-negative real-valued vectors in which we want to solve a constraining problem on the values of the PSA measurement. A particular constraining problem, is represented by thresholding, i.e. finding couples $(\mathbf{x_i},\mathbf{x_j}) \in \mathcal{A} \subseteq R_0^{+N}$ such that $S_{m,n}(\mathbf{x_i},\mathbf{x_j}) \geq \alpha$. A way with which this problem can be solved is constituted by analysing the adjacency matrix

$$\mathbf{A} = \mathbf{A}_{\cos} \bullet \mathcal{F}^{n,\varDelta} = A_{\cos} \bullet \mathbf{\Phi}^{1-n-\varDelta} \bullet \mathbf{\Phi}^{1-n^T} \tag{9}$$

where $a_{ij} = S_{m,n}(\mathbf{x_i},\mathbf{x_j})$, $\bullet$ is the Hadamard product, $\mathbf{A}_{\cos}$ is the Cosine similarity adjacency matrix, and $\phi_{ij} = \|\mathbf{x_i}\|$. This re-formulation allows for the interpretation that an adjacency matrix corresponding to a PSA $S_{n+\varDelta,n}(\mathbf{x_i},\mathbf{x_j})$ is a *distortion* of the Cosine similarity adjacency matrix $\mathbf{A}$ obtained by Hadamard-multiplying $\mathbf{A}_{\cos}$ by the term $\mathbf{\Phi}^{1-n-\varDelta} \bullet \mathbf{\Phi}^{1-n^T}$. We observe that the matrix $\mathbf{\Phi}$ is a static feature for $\mathcal{A}$, i.e. it can be calculated once for all, and that the level of distortion of $\mathbf{A}_{\cos}$ depends only on n and $\varDelta$. To give empirical examples of these features, let us consider as $\mathcal{A}$ a set of randomly generated vectors in the space $[0,1] \times [0,1]$ and let us add the element $\mathbf{x}^* = (1/2,1/2) \in \mathcal{A}$. We then calculate the matrix $\mathbf{A}$ of Eq. (9) and find the following subsets of $\mathcal{A}$:

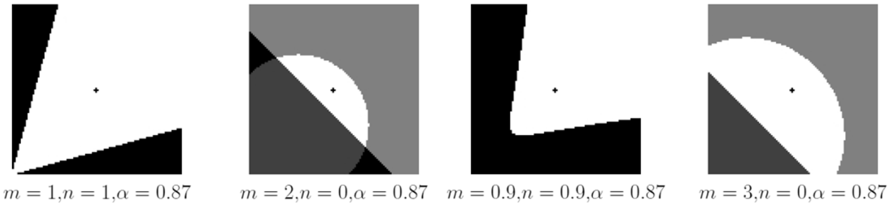
$$Eq_{x^*} = \{x \in \mathcal{A} : S_{m,n}(x^*,x) \geq \alpha \wedge S_{m,n}(x,x^*) \geq \alpha\} \tag{10}$$

$$Dis_{x^*} = \{x \in \mathcal{A} : S_{m,n}(x^*,x) < \alpha \wedge S_{m,n}(x,x^*) < \alpha\} \tag{11}$$

$$En_{x^*} = \{x \in \mathcal{A} : S_{m,n}(x^*,x) \geq \alpha \wedge S_{m,n}(x,x^*) < \alpha\} \tag{12}$$

$$Ien_{x^*} = \{x \in \mathcal{A} : S_{m,n}(x^*,x) < \alpha \wedge S_{m,n}(x,x^*) \geq \alpha\} \tag{13}$$

Eqs. (10) and (11) denote, respectively, syntactic equivalence and syntactic disjointness between vectors $\mathbf{x}^*$ and $\mathbf{x}$. Eq. (12) denotes a syntactic entailment from $\mathbf{x}^*$ to $\mathbf{x}$. Eq. (13) denotes a (inverse) syntactic entailment to $\mathbf{x}^*$ from $\mathbf{x}$ (see [6] for further detail). Examples of the results for four possible configurations of m, n, with $\alpha = 0.87$ are shown in Fig. 1, where subsets Eq_{x^*}, Dis_{x^*}, En_{x^*}, and Ien_{x^*} are rendered in white, black, light grey, and dark grey respectively.



$m=1,n=1,\alpha=0.87$ $m=2,n=0,\alpha=0.87$ $m=0.9,n=0.9,\alpha=0.87$ $m=3,n=0,\alpha=0.87$

Fig. 1. Examples of space regions for PSA with different combinations of m, n and α. The white (black) area denotes a syntactic equivalence (disjointness) of two vectors $\mathbf{x}, \mathbf{x}^*$. The light (dark) gray area denotes a syntactic entailment from (to) vector $\mathbf{x}$ to (from) vector $\mathbf{x}^*$. Note that, only either equivalence or disjointness relations are representable using a symmetric function (first and third sub-figure).

2.2 Illustrative Examples

A particular case of Eq. (1) is when $m = 2$ and $n = 0$. In this situation the PSA measure has some interesting properties w.r.t. plain Cosine similarity, which will be illustrated by some examples in this Section. Empirical investigations demonstrated the effectiveness of choosing $S_{2,0}(\cdot)$ as affinity function for real-world clustering problems (see §4). Let us consider the following vectors:

$$\mathbf{a} = (0\ 1\ 1\ 1\ 0\ 0\ 0\ 0) \quad \mathbf{b} = (0\ 1\ 1\ 1\ 0\ 0\ 1\ 1) \quad \mathbf{c} = (0\ 0.5\ 0.5\ 0.5\ 0\ 0\ 0\ 0) = \frac{\mathbf{a}}{2}$$

for which it results $S_{1,1}(\mathbf{a}, \mathbf{b}) = \cos(\mathbf{a}, \mathbf{b}) \approxeq 0.77$, $S_{2,0}(\mathbf{a}, \mathbf{b}) = 1$, $S_{2,0}(\mathbf{b}, \mathbf{a}) = 0.6$, $S_{1,1}(\mathbf{c}, \mathbf{b}) = \cos(\mathbf{c}, \mathbf{b}) \approxeq 0.77$, $S_{2,0}(\mathbf{c}, \mathbf{b}) = 2$, $S_{2,0}(\mathbf{b}, \mathbf{c}) = 0.3$, $S_{2,0}(\mathbf{c}, \mathbf{a}) = 2$ and $S_{2,0}(\mathbf{a}, \mathbf{c}) = 0.5$. It can be observed that $S_{1,1}(\mathbf{c}, \mathbf{b}) = S_{1,1}(\mathbf{a}, \mathbf{b})$, $S_{2,0}(\mathbf{a}, \mathbf{c}) = S_{2,0}(\mathbf{a}, \frac{\mathbf{a}}{2}) = \frac{1}{2} S_{2,0}(\mathbf{a}, \mathbf{a})$ and $S_{2,0}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} S_{2,0}(\mathbf{c}, \mathbf{b})$, as defined by the analytical properties of the PSA measure.

These examples enlighten the following qualitative properties of the pseudo-semantic affinity measurement introduced in this research:

- The pseudo-semantic affinity measure accounts compactly for the asymmetric relationship existing among vectors in terms of angle and size at the same time;
- as such, and differently from Cosine similarity, the pseudo-semantic affinity is sensible to vector scaling, and accounts proportionally to the scaling coefficient.

A natural synthesis of the two above example brings to state that vector $\mathbf{a}$ is *completely included* in vector $\mathbf{b}$, and that vector $\mathbf{c}$ is *doubly included* in vector $\mathbf{b}$. Furthermore we can say that vector $\mathbf{b}$ is *included by* $\frac{3}{5}$ in vector $\mathbf{a}$. Intuitively, it means that $S_{2,0}(\mathbf{a}, \mathbf{b})$ accounts for the extent to which the vector $\mathbf{a}$ is explained by the vector $\mathbf{b}$.

As a final example, let us consider a vector $\mathbf{e}$ defined as $\mathbf{e} = \mathbf{a} + \epsilon$ with ϵ orthogonal both to $\mathbf{a}$ and $\mathbf{b}$ and $\rho = \frac{\|\epsilon\|}{\|\mathbf{a}\|} \ll 1$, i.e. a vector representing an additive bi-orthogonal noise to $\mathbf{a}$. This example models situations in which there are random errors in the calculation of feature vectors due to uncorrelated noisy input, e.g. spurious terms included in a textual feature vector by an imperfect HTML text extraction engine, or wrong words in the transcription of spoken content. The PSA measures between **e** and **b** are:

$$\cos(\mathbf{e}, \mathbf{b}) = \cos(\mathbf{b}, \mathbf{e}) \approx \cos(\mathbf{a}, \mathbf{b}) \left(1 - \frac{1}{2}\rho^2\right)$$

$$S_{2,0}(\mathbf{e}, \mathbf{b}) \approx S_{2,0}(\mathbf{a}, \mathbf{b}) \left(1 - \rho^2\right) \text{ and } S_{2,0}(\mathbf{b}, \mathbf{e}) = S_{2,0}(\mathbf{b}, \mathbf{a})$$

These results enlighten an important feature distinguishing the pseudo-semantic affinity measurement from Cosine similarity: addition of bi-orthogonal noise to a vector *always* affects Cosine similarity, while it affects semantic affinity only partially. In other words, the *degree of inclusion* of a vector **a** when perturbed with by bi-orthogonal noise ϵ in another vector **b** is affected, but the opposite does not hold, i.e. the *degree of inclusion* of a vector **b** in a vector **a** is not affected by bi-orthogonal noise added to **a**.

3 Co-clustering Algorithm

Let $\mathit{\Pi} = \{\pi_i\}_{i=1}^m$ and $\mathcal{B} = \{\beta_j\}_{j=1}^n$ be two input spaces of data sets, for which a metric in the hybrid output space $\mathcal{H} = \mathit{\Pi} \cup \mathcal{B}$ is not defined. Our co-clustering algorithm uses the PSA measure to aggregate the objects from the input spaces in the output space $\mathcal{H}$. Figure 2 shows all the procedural blocks involved in the framework. The data from the ancillary space ($\mathcal{B}$) are used as a *dynamic dictionary* with which the data from the primary space ($\mathit{\Pi}$) are represented. Each aggregation $\gamma_i^* \in D^*(\alpha)$ consists of two components:

- A subset of items of $\mathit{\Pi}$ which share common attributes in $\mathcal{B}$, as given by a PSA measurement $S_{m,n}(\pi_a, \pi_b)$, $\pi_a, \pi_b \in \mathit{\Pi}$, which operates on the cross-relevance between items of $\mathit{\Pi}$ and items of $\mathcal{B}$ condensed by the transformation $\mathcal{T}$ into the space $\mathcal{A}$. It is possible to demonstrate that the PSA is a pseudo-quasi-metric and thus can be used to compare the heterogeneous data objects in $\mathcal{A}$;
- The attributes of the dictionary $\mathcal{B}$ which are relevant to each or all depending on the chosen criterion for multimodal aggregation (i.e. either intersection or union) aggregated elements of γ_i;

The following subsections describes all the constituent parts of the framework.

3.1 Relevance Cross-Projection

Let $\mathcal{T} : \mathit{\Pi} \times \mathcal{B} \rightarrow [0, 1]$ be a linking function such that $\mathcal{T}(\pi_i, \beta_j)$ tends to 1 if $\beta_j \in \mathcal{B}$ is somewhat relevant to $\pi_i \in \mathit{\Pi}$, and $\mathcal{T}(\pi_i, \beta_j)$ tends to 0 if $\beta_j \in \mathcal{B}$ is not relevant to $\pi_k \in \mathit{\Pi}$. As an example, $\mathcal{T}(\cdot)$ could be the score function that project the data from the primary space $\mathit{\Pi}$ to the data of the ancillary space $\mathcal{B}$. The relations between items from the two spaces can be thus expressed by the cross-projection matrix **T**, whose elements are $t_{ij} = \mathcal{T}(\pi_i, \beta_j)$.

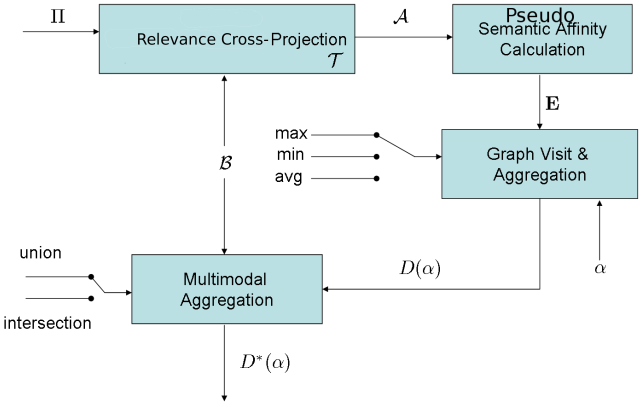


Fig. 2. Parts of the co-clustering framework.

3.2 Pseudo-Semantic Affinity Calculation

From the matrix **T** we evaluate the affinity between all the couples of the primary space objects. Formally, let π_a, π_b be two objects from the primary space, represented by the cross-projection vectors $\mathbf{t_a} = \left(\mathcal{T}(\pi_a, \beta_j)_{j=1}^n\right)$, $\mathbf{t_b} = \left(\mathcal{T}(\pi_b, \beta_j)_{j=1}^n\right)$, $\mathbf{t_a}, \mathbf{t_b} \in \mathbf{T}$. Then the PSA of order m, n from $\mathbf{t_a}$ to $\mathbf{t_b}$ is evaluated according to Eq. (1), and its value is stored in the affinity matrix $\mathbf{E} = (e_{ab})$, where

$$e_{ab} = \begin{cases} 1, \text{ if } a = b \\ S(\mathbf{t_a}, \mathbf{t_b}), \text{ if } a \neq b \text{ and } S(\mathbf{t_a}, \mathbf{t_b}) \geq \alpha \\ 0, \text{ otherwise.} \end{cases} \tag{14}$$

3.3 Graph Visit and Multimodal Aggregation

The affinity matrix **E** can be interpreted as the adjacency matrix of a directed disconnected graph $G = (V, E)$ whose vertices are the primary space objects and whose edges are the affinities among them. The co-clustering problem can be thus reduced to a graph visiting problem. For that, we first *symmetrise* **E**, by considering either connected or not each couple of nodes. then we use the depth first search (DFS) algorithm, obtaining the set $D(\alpha) = \{g_1, g_2, \ldots, g_K\}$ so that $\bigcup_i g_i = G$ and $\forall i \neq j,\ g_j \bigcap g_i = \emptyset$ of disconnected sub-graphs included in G. Since the DFS algorithm is valid for not oriented graphs, first we generate a not oriented representation of the original oriented graph using one of the following criteria:

- Maximum affinity: vertices a, b are linked if $\max\left(S(\mathbf{t_a}, \mathbf{t_b}), S(\mathbf{t_b}, \mathbf{t_a})\right) > \alpha$
- Average affinity: vertices a, b are linked if $\frac{1}{2}\left(S(\mathbf{t_a}, \mathbf{t_b}) + S(\mathbf{t_b}, \mathbf{t_a})\right) > \alpha$
- Minimum affinity: vertices a, b are linked if $\min\left(S(\mathbf{t_a}, \mathbf{t_b}), S(\mathbf{t_b}, \mathbf{t_a})\right) > \alpha$

Given $D(\alpha)$ we can finally build the set of multimodal aggregations $D(\alpha)^* = \{\gamma_1^*, \ldots, \gamma_{|D(\alpha)|}^*\} \subseteq 2^{\mathcal{H}}$ by collecting the elements of $\mathcal{B}$ which are relevant to each $\gamma_i \in D(\alpha)$. Letting $K = |D(\alpha)|$, this can be done according to one of the following criteria.

Union criterion

$$\forall i : \gamma_i^* = \gamma_i \cup B_i, \quad i = 1, \ldots, K \tag{15}$$

$$B_i = \cup_{j=1}^{|\gamma_i|} \beta_{ij} \tag{16}$$

$$\beta_{ij} = \{b \in \mathcal{B} : R(\pi_{ij}, b) > \eta\} \ , \tag{17}$$

where η is a parametric threshold. Following this criterion, the function of $D(\alpha)^*$ is that of integrating the partition $D(\alpha)$ with *all* the relevant elements of $\mathcal{B}$. Notice that $D(\alpha)^*$ is not in general a partition of $\mathcal{H} = \Pi \cup \mathcal{B}$, because elements of $\mathcal{B}$ may be relevant to elements of Π belonging to different elements of $D(\alpha)$, and because some elements of $\mathcal{B}$ may be not relevant to any element of Π.

Intersection criterion

$$\forall i : \gamma_i^* = \gamma_i \cup B_i, \quad i = 1, \ldots, K \tag{18}$$

$$B_i = \cup_{j=1}^{|\gamma_i|} \beta_{ij} \tag{19}$$

$$\beta_{ij} = \{b \in \mathcal{B} : \forall k : R(\pi_{ik}, b) > \eta\} \ , \tag{20}$$

Following this criterion, the function of $D(\alpha)^*$ is that of integrating the partition $D(\alpha)$ with the elements of $\mathcal{B}$ that are relevant to *each* element of γ_i. Notice that neither in this case $D(\alpha)^*$ is a partition of $\mathcal{H} = \Pi \cup \mathcal{B}$.

4 Experimental Results

This section provides empirical evidence to demonstrate the benefits of our co-clustering framework and algorithm. In particular, we show that the co-clustering approach based on PSA measurement overcomes clustering methods based on symmetric similarity measures. For this purpose, we applied the algorithm to the real case of aggregating news content from the Web and TV broadcasts, i.e. the primary space Π and the ancillary space $\mathcal{B}$ defined in §3, respectively. For the evaluations we set up a pool of 25 users, taken from the employers of our organisation, who were unaware of the rationales of the framework.

We randomly collected a set of 2, 155 news articles taken from 16 different newspapers and press agencies websites. The articles were randomly chosen in order to cover a broad range of subjects such as politics, sports, entertainment, current events and foreign affairs. Two different feature extraction and clustering strategies were applied to these data.

The first strategy adopted the common term frequency-inverse document frequency (tf/idf) weighting-based vector model and Cosine similarity. Each Web article was represented by the tf/idf values for the terms occurring in the document. Clustering was performed using Eq. (8) as distance metric, where we set $n = 1$ and $\Delta = 0$. The minimum affinity criteria for different values of α was used for determining whether two graph vertices were connected or not.

In the second strategy we first performed part-of-speech tagging , selecting for each article the set of words tagged as nouns, proper nouns or adjectives. These word sets were then matched with the speech content of about $24,000$ TV news stories, resulting from the automatic segmentation of $3,670$ newscast programmes. Each Web article was thus represented by a cross-projection vector counting the relevance scores of the text documents obtained by the automatic transcription of each TV story's speech content w.r.t. the set of words extracted from the article. This procedure practically implements the relevance cross-projection described in §3.1. Further detail on the overall process can be found in [6]. For each couple of Web articles a, b the pseudo-semantic affinities $S_{m,n}(a, b)$ and $S_{m,n}(b, a)$ were calculated across several combinations of m, n. Clustering was performed applying the maximum affinity criteria for different values of α and using the union criterion with $\eta = 0.55$.

We evaluated the effectiveness of our framework by measuring the clustering cohesion (Ω), the clustering coverage (Ψ) and the clustering efficiency ($\Lambda = \frac{2\Omega\Psi}{\Omega+\Psi}$). The clustering cohesion can be related to the precision of the algorithm and is calculated as the ratio of objects in a cluster that are relevant to the concept expressed by the cluster and the total number of objects included in the same cluster averaged on the set of detected clusters. The concept expressed by the cluster is the concept that the user assumed as such during the assessment of a specific cluster. This implicitly means that performance should not empirically be lower than 0.5, since it is expected that in a situation in which the two halves of a cluster are equally distributed between two concepts the user would choose one of the two as the reference concept and discard the other. The clustering coverage can be related to the recall of the algorithm and is calculated as the ratio between the average number of objects included in a cluster with cardinality greater than 1 (Avg. Card. * Tot. Clusters) and the total number of objects to be clustered. The clustering efficiency can be related to the F-measure accounting for the balance between cohesion of the detected clusters and their expected cardinality.

Table 1 shows the performance results obtained for the tf/idf-based clustering strategy. Table 2 shows the performance results obtained for the relevance cross-projection-based co-clustering strategy using different symmetric combinations of the PSA measure. Table 3 compares these two clustering approaches in terms of clustering efficiency. Observe that the cross-projection-based strategy always outperforms the tf/idf-based strategy. Table 4 shows the performance results obtained for the relevance cross-projection-based co-clustering strategy using generic combinations of the asymmetric PSA measure. Results indicate that best performers are not always the same. For example for $\alpha = 0.5$ best cohesion is achieved for $m = 7$ and $n = 3$, while for $\alpha = 0.95$ it is $m = 5$ and $n = 0$. However, it can be observed that under some combinations of m and n, the algorithm always achieves good results independently from the chosen value of α. In the absence of the concrete possibility of full ground-truth testing of all possible combinations, these or similar global observations may help in finding the proper choice of PSA parameters for a given information retrieval system. For example, Table 4 shows that the combination $(m, n, \alpha) = (2, 0, 0.85)$ provides better balance between the clustering cohesion and coverage for the studied case. Though these findings are not enough to identify clear conditions and criteria under which to select the proper parameter combination for an information retrieval system, we think that at least they clearly show that the generalisation towards the PSA domain is not useless nor trivial to resolve, and as such needs further thorough investigation.

Table 1. Performance of the clustering procedure using the tf/idf vector space model and the Cosine similarity, i.e. $n = 1$ and $\Delta = 0$ in Eq. (8).

Avg. Card.	Max. Card.	Tot. Clusters	α	Ω	Ψ	Avg. Card.	Max. Card.	Tot. Clusters	α	Ω	Ψ
2.00	2	24	0.98	1	0.02	2.61	7	41	0.95	0.85	0.05
5.56	48	55	0.87	0.53	0.14	6.34	78	104	0.85	0.36	0.31

Table 2. Performance of the co-clustering procedure using the relevance cross-projection vector space model and the PSA-based symmetric function, i.e. $m = n$ in Eq. (1). Note that for $m = n = 1$ the PSA function is reduced to the Cosine function.

Avg. Card.	Max. Card.	Tot. Clusters	m	n	α	Ω	Ψ	Avg. Card.	Max. Card.	Tot. Clusters	m	n	α	Ω	Ψ
2.49	7	291	1	1	0.98	0.94	0.34	2.47	7	309	1	1	0.95	0.88	0.35
2.53	7	342	1	1	0.87	0.90	0.40	2.18	4	22	2	2	0.87	0.83	0.02
2.0	2	2	5	5	0.87	0.5	0.002	2.0	2	6	3	3	0.87	0.5	0.006
2.08	3	13	5	5	0.85	1	0.01	2.83	9	419	1	1	0.85	0.91	0.55
2.28	4	64	2	2	0.85	0.89	0.07	2.17	4	23	3	3	0.85	0.83	0.02
2.24	4	50	4	4	0.5	0.93	0.05	2.29	4	86	3	3	0.5	0.90	0.09
2.50	7	177	2	2	0.5	0.89	0.21	4.04	27	453	1	1	0.5	0.79	0.85

5 Graph-Based Interaction Tools

In order to provide interactive visualisation of the generated multimodal aggregations for news retrieval, browsing and editing, we developed a browsing interface that enables users to track down the structural properties of the selected aggregation and, at the same time, navigate and manage its contents. This functionality is inspired by hyperlink networks, e.g. social and citation networks, Internet

Table 3. Comparison of clustering efficiency between the tf/idf-based clustering procedure and the relevance cross-projection-based co-clustering procedure using the Cosine similarity as distance metric.

Method	α	Efficiency	Method	α	Efficiency
tf/idf	0.98	0.04	relevance cross-projection	0.98	0.5
tf/idf	0.95	0.09	relevance cross-projection	0.95	0.5
tf/idf	0.87	0.22	relevance cross-projection	0.87	0.55
tf/idf	0.85	0.33	relevance cross-projection	0.85	0.69
tf/idf	0.5	0.23	relevance cross-projection	0.5	0.82

communications, and graph theory. The structure of each multimodal aggregation is represented as an oriented graph where the included Web articles are the nodes and the relationships among them are the edges, reflecting the asymmetric structure obtained by applying the PSA-based co-clustering algorithm. The graphs are browsable through a Java Web interface. Users can select the visualisation layout to apply for graph display (default is Fruchterman-Reingold layout - FRL). The nodes are labelled according to the titles of the Web articles, and are sized based to their representativeness (i.e., in-degree) w.r.t. the main topic of the aggregation, so that more representative nodes are drawn with larger shapes and bolder colors. Moving on one node shows the information related to the node, in terms of node's title, in-degree and out-degree. Similarly, moving on one edge shows the edge weight. Starting from one node, the user can select, traverse and manipulate all its context/detail nodes. Context nodes of a node $\mathbf{v}$ are those nodes $\mathbf{z}$ for which $S(\mathbf{v}, \mathbf{z}) \geq \alpha$, while detail nodes are those nodes w for which $S(\mathbf{w}, \mathbf{v}) \geq \alpha$. This functionality allows users to find semantically meaningful paths between the news items of the aggregation.

The use of the FRL graph layout algorithm allows for intuitive visual-aided data navigation and manipulation. The assumption is that the graph's topology can be used to describe the underlying concepts of the aggregation. Cohesive aggregations would be modelled by dense graphs. On the other hand, less cohesive aggregations would be modelled by spread graphs. Figures 3 and 4 show two examples. The former illustrates how the system is able to visually distinguish two subtopics in the context of a larger topic. The overall topic is about the summer traffic jams. The two subtopics (one of which is highlighted in yellow) are aggregations reporting of the same main topic (traffic jams) over two different consecutive weeks, marked respectively as "red" and "black" by the Italian Transport Ministry. The latter expresses still the same concept in the case of January 2009 bad weather period. However in this case the highlighted part is referring to the disasters happened in France and Spain (foreign countries) while the remain part is referring to Italian casualties.

Table 4. Performance of the co-clustering procedure using the relevance cross-projection vector space model and the PSA-based asymmetric function, i.e. $m \neq n$ in Eq. (1).

Avg. Card.	Max. Card.	Tot. Clusters	m	n	α	Ω	Ψ	Avg. Card.	Max. Card.	Tot. Clusters	m	n	α	Ω	Ψ
2.7	4	10	4	0	0.98	1	0.01	2.58	4	12	3	0	0.98	1	0.01
2.32	5	22	3	1	0.98	0.96	0.02	2.57	4	21	2	0	0.98	0.94	0.03
2.56	6	39	2	1	0.98	0.89	0.05	2.69	4	16	5	0	0.95	1	0.02
2.51	6	47	3	1	0.95	0.95	0.05	2.47	6	75	2	1	0.95	0.90	0.09
2.65	5	26	3	0	0.95	0.90	0.03	2.67	6	39	2	0	0.95	0.80	0.05
2.0	2	7	6	0	0.87	1	0.01	2.58	7	138	2	1	0.87	0.95	0.17
3.34	7	106	2	0	0.87	0.94	0.16	2.84	7	62	3	0	0.87	0.93	0.08
2.54	7	99	3	1	0.87	0.80	0.12	2.08	3	13	6	3	0.85	1	0.01
3.12	16	144	3	0	0.85	0.96	0.21	2.69	7	142	3	1	0.85	0.94	0.18
2.68	7	200	2	1	0.85	0.93	0.25	**4.56**	**22**	**219**	**2**	**0**	**0.85**	**0.95**	**0.47**

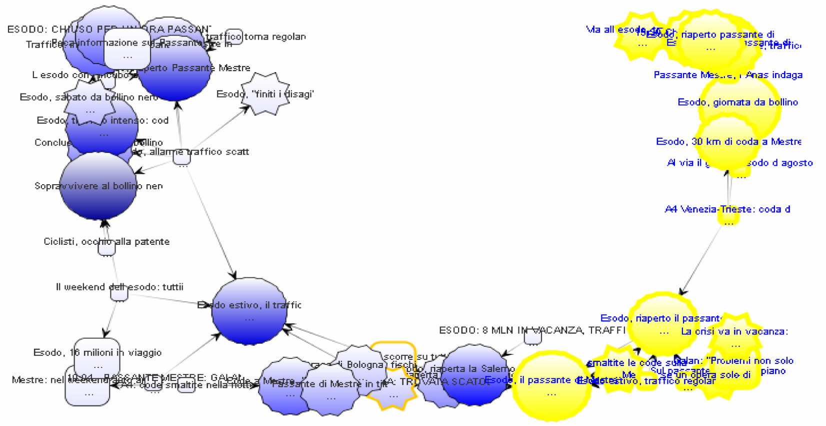
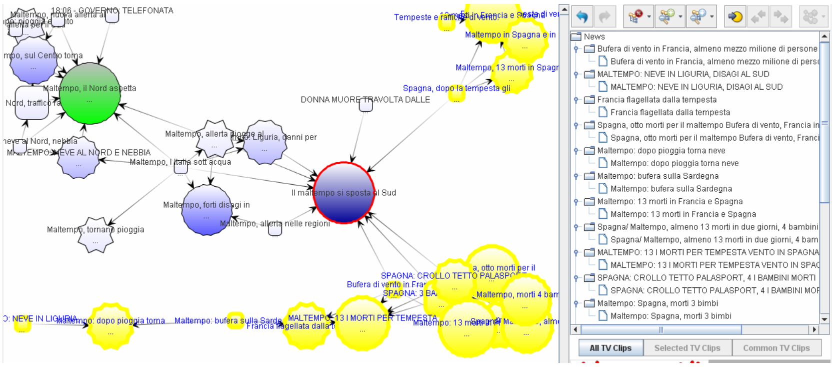

Esodo, "finiti i disagi"
Sopravvivere al bollino ner
Ciclisti, occhio alla patente
Il weekend dell esodo: tutti
Esodo estivo, il traffic
Esodo, 16 milioni in viaggio
ESODO: 8 MLN IN VACANZA, TRAFFI
Esodo, giornata da bollino
Esodo, 30 km di coda a Mestre
A4 Venezia-Trieste: coda d
Esodo, riaperto il passant
La crisi va in vacanza:
Passante di Mestre in tilt

Fig. 3. Visual-aided graph-based news navigation - example 1.


18:06 - GOVERNO: TELEFONATA
Maltempo, il Nord aspetta
DONNA MUORE TRAVOLTA DALLE
Spagna, dopo la tempesta gli
Maltempo, 13 morti in Spagn
Maltempo in Spagna e in
Il maltempo si sposta al Sud
Maltempo, forti disagi in
Maltempo, allerta nelle regioni
Maltempo, tornano pioggia
Maltempo, l'Italia sott acqua
SPAGNA: CROLLO TETTO PALASPORT
Maltempo: dopo pioggia torna
Maltempo: bufera sulla Sardeg
MALTEMPO: 13 I MORTI PER TEMPESTA
All TV Clips
Selected TV Clips
Common TV Clips

News
- Bufera di vento in Francia, almeno mezzo milione di persone
  - Bufera di vento in Francia, almeno mezzo milione di perso
- MALTEMPO: NEVE IN LIGURIA, DISAGI AL SUD
  - MALTEMPO: NEVE IN LIGURIA, DISAGI AL SUD
- Francia flagellata dalla tempesta
  - Francia flagellata dalla tempesta
- Spagna, otto morti per il maltempo Bufera di vento, Francia in
  - Spagna, otto morti per il maltempo Bufera di vento, Franci
- Maltempo: dopo pioggia torna neve
  - Maltempo: dopo pioggia torna neve
- Maltempo: bufera sulla Sardegna
  - Maltempo: bufera sulla Sardegna
- Maltempo: 13 morti in Francia e Spagna
  - Maltempo: 13 morti in Francia e Spagna
- Spagna/ Maltempo, almeno 13 morti in due giorni, 4 bambini
  - Spagna/ Maltempo, almeno 13 morti in due giorni, 4 bam
- MALTEMPO: 13 I MORTI PER TEMPESTA VENTO IN SPAGNA
  - MALTEMPO: 13 I MORTI PER TEMPESTA VENTO IN SPAC
- SPAGNA: CROLLO TETTO PALASPORT, 4 I BAMBINI MORTI
  - SPAGNA: CROLLO TETTO PALASPORT, 4 I BAMBINI MOR
- Maltempo: Spagna, morti 3 bimbi
  - Maltempo: Spagna, morti 3 bimbi

Fig. 4. Visual-aided graph-based news navigation - example 2.

6 Conclusions and Future Work

This paper presented a family of pseudo-semantic affinity (PSA) functions for modelling the relationships between heterogeneous data objects. A co-clustering framework based on these functions was also presented. The motivation behind the use of this framework is its capability of representing hierarchical relationships among data of different nature, which are not directly derivable by using one-way clustering techniques (e.g. tf/idf vector space model and Cosine similarity). The experiments demonstrate that PSA measure outperforms the Cosine similarity in terms of clusters' cohesion and coverage, thus providing better precision and recall. It was also demonstrated that the PSA is only partially affected by noisy data. Future work will explore how to adaptively select the best combination of PSA parameters for the optimisation of information retrieval processes.

References

1. I. S. Dhillon. Co-clustering documents and words using bipartite spectral graph partitioning. In *KDD '01: Proc. of the 7th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 269–274, 2001.
2. C. Domeniconi and D. Gunopulos and S. Ma and B. Yan and M. Al-Razgan and D. Papadopoulos Locally Adaptive Metrics for Clustering High Dimensional Data. *Data Mining and Knowledge Discovery Journal*, 14(1):63–97, 2007.
3. C. Domeniconi and J. Peng and B. Yan Composite Kernels for Semi-supervised Clustering. *Knowledge and Information Systems*, 1–18, 2010.
4. V. Kashyap and A. Sheth. Semantic and schematic similarities between database objects: a context-based approach. *The VLDB Journal*, 5(4):276–304, 1996.
5. B. Long, Z. Zhang, and P. S. Yu. Co-clustering by block value decomposition. In *KDD '05: Proc. of the 11th ACM SIGKDD Intl. Conf. on Knowledge discovery and data mining*, pages 635–640, 2005.
6. A. Messina and M. Montagnuolo. A generalised cross-modal clustering method applied to multimedia news semantic indexing and retrieval. In *WWW '09: Proc. of the 18th Intl. Conf. on World wide web*, pages 321–330, 2009.
7. M. Rege, M. Dong, and F. Fotouhi. Co-clustering image features and semantic concepts. In *Proc. of the Intl. Conf. on Image Processing*, pages 137–140, 2006.