Multi-source shared nearest neighbours for multi-modal image clustering

Amel Hamzaoui, Alexis Joly, Nozha Boujemaa

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE

Multi-source shared nearest neighbours for multi-modal image clustering

Amel HAMZAOUI — Alexis JOLY — Nozha BOUJEMAA

N° 0000

Juillet 2010

Thème COG

ISSN 0249-6399 ISRN INRIA/RR--0000--FR+ENG

INRIA
ROCQUENCOURT

Multi-source shared nearest neighbours for multi-modal image clustering

Amel HAMZAOUI[†], Alexis JOLY [†], Nozha BOUJEMAA [†]

Thème COG — Systèmes cognitifs
Projet IMEDIA

Rapport de recherche n° 0000 — Juillet 2010 — 14 pages

Abstract: Shared Nearest Neighbours (SNN) techniques are well known to overcome several shortcomings of traditional clustering approaches, notably high dimensionality and metric limitations. However, previous methods were limited to a single information source whereas such methods appear to be very well suited for heterogeneous data, typically in multi-modal contexts. In this paper, we introduce a new **multi-source** shared neighbours scheme applied to multi-modal image clustering. We first extend existing SNN-based similarity measures to the case of multiple sources and we introduce an original automatic source selection step when building candidate clusters. The key point is that each resulting cluster is built with its own optimal subset of modalities which improves the robustness to noisy or outlier information sources. We experiment our method in the scope of multimodal image search results clustering and show its effectiveness using both synthetic and real data involving different visual and textual information sources and several datasets of the literature.

Key-words: Multi Source, Clustering, Search results, shared neighbors, multimodality.

[†] INRIA-Rocquencourt, team-project IMEDIA, BP 105, F-78153 Le Chesnay Cedex (France); e-mails: Amel.Hamzaoui@inria.fr, Alexis.Joly@inria.fr,Nozha.Boujemaa@inria.fr.

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France)
Téléphone : +33 1 39 63 55 11 — Télécopie : +33 1 39 63 53 30

Structuration multi-source basée sur les voisins partagés appliquée aux images multi-modales.

Résumé :
Les techniques basées sur l'information des voisins partagés sont bien connues pour surmonter plusieurs lacunes des méthodes traditionnelles de regroupement, celles liés aussi à la grande dimension et les limites des métriques. Cependant, les méthodes précédentes étaient limitées à une seule source d'information alors qu'elles semblent être très bien adaptées pour des données hétérogènes, généralement dans un contexte multi-modal.
Dans cet article, nous introduisons une nouvelle () approche multi-source appliquée à une méthode de regroupement basée sur l'information des voisins partagés. Nous avons d'abord étendu les mesures de similarité dans un cas de multiples sources et nous avons introduit une étape originale qui permet la sélection automatique des sources lors de la construction des groupes candidats.
L'originalité de la méthode est que chaque groupe qui en résulte est construit avec ses propres sous-ensembles de modalités qui lui sont optimales et qui améliore la robustesse face aux sources d'information bruitées ou aberrantes.
Nous avons expérimenté notre méthode dans le cadre de la structuration de résultats de recherche d'images de façon multi-modale et nous avons montré son efficacité en utilisant des données synthétiques et d'autres réelles impliquant différentes sources d'information visuelle et textuelle.

Mots-clés : Multi-source, structuration, rèsultat de recherche, Voisins partagés, Multi-modalités

Table of contents

1 Introduction

Unsupervised data clustering remains a crucial step of many recent multimedia retrieval approaches, e.g. web objects and events mining [1], search results clustering [2] or visual query suggestion [3]. However, the performance and applicability of many classical data clustering approaches often force particular choices of data representation and similarity measures. Some methods, such as k-means and its variants [4], require the use of L_p metrics or other specific measures of data similarity; others, such as the hierarchical methods BIRCH [5] and CURE [6], pay a prohibitive computational cost when the representational dimension is high, due to their reliance on data structures that depend heavily upon the data representation. Such assumptions are particularly problematic in a multimedia context that usually involves heterogeneous data and similarity measures.
An interesting alternative approach to clustering that requires only comparative tests of similarity values is the use of so-called shared-neighbor information [7],[8],[9],[10],[11] and [12]. Here, two items are considered to be well-associated not by virtue of their pairwise similarity value, but by the degree to which their neighborhoods resemble one another. Even in heterogeneous contexts in which underlying features and similarity values do not have a straightforward unique interpretation, two items having a high proportion of neighbors in common can be assign to the same group. Such methods are known to overcome several shortcomings of traditional clustering approaches: they do not suffer from dimensionality curse, they are robust to noisy data, they do not need to initially fix the number of clusters, and, last but not least, they do not require any explicit knowledge of the nature or representation of the data. These properties make them widely generic for any multimedia mining or structuring purpose, whatever the targeted objects and the required similarity measures.
Shared Nearest Neighbors (**SNN**) methods thus appear to be ideally suited to **multi-modal** clustering. Because they are based on conexity information only and not on densities or metrics in some feature spaces, heterogeneous information sources can be embedded identically and easily compared or fused. But surprisingly, very few works in the literature have been addressing this Multi-source SNN aspect (cf. section 2); this is the main contribution of our paper. We introduce a new generic **multi-source** SNN framework including new multi-source intra-set and inter-set significance measures for arbitrary object sets and information sources. The main originality of our approach is that we introduce an information source selection step in the computation of these measures thanks to an a contrario standardization of the sum of the individual SNN scores. In addition to a usual conexity score, any arbitrary object set is thus associated with its own optimal subset of modalities maximizing the normalized multi-source significance measure. As shown in the experiments, this source selection step makes our approach widely robust to the presence of locally outlier sources, i.e. sources producing non relevant nearest neighbors (e.g. close to random) for some input objects or clusters.
The rest of the paper is organized as follow: Previous works related to SNN are reviewed in section 2. Our new Multi-source SNN approach is introduced in section 3 and the overall clustering framework in section 4 . Experiments are reported in section 5.

2 Related Works

The origins of the use of neighborhood information for clustering can be traced to the shared-neighbor merge criterion of Jarvis and Patrick [7] used in agglomerative clustering. The criterion states that two clusters can be merged if they contain equal-sized sub clusters A and B such that $\mid A \cap B \mid \geq m\dot{k}$, where k is the size of A and B, and $0 < m < 1$ is a fixed merge threshold parameter. As is often the case with the agglomerative clustering, this merge criterion can result in clusters composed of long chains of sub clusters in which the items at one end of the chain is very different to those of the other end.

One influential shared-neighbor clustering method that improved above Jarvis-Patrick work is the hierarchical algorithm ROCK [11]. ROCK makes the merge criterion depending on the degree of overlap between neighbours sets of clusters items. Furthermore it avoids the chaining problem by assessing linkages over all pairs of clusters items. However, as Jarvis-Patrick work, ROCK requires to fix the appropriate neighborhood size k, which appears to be a crucial and sensitive parameter of shared neighbors methods.
Another noticeable SN-clustering derived from the well-known density based clustering DBSCAN [13] is SNN [10]. Whereas DBCSAN works by detecting seed points in regions of high density in the feature space, SNN introduces a new SN-based density measure estimated by the sum of the intersections between the k-neighborhood of the candidate seed and the k-neighborhoods of its neighbors. Focusing on connectivity rather than feature space density, SNN has the advantage to detect locally dense clusters even they are globally of low density compared to other clusters. The drawback of the method, however, is that it still requires a user-supplied neighborhood size k and a threshold on the density measure to identify the seed points.
The PatClust algorithm, proposed by [14], can be seen as an extension of SNN allowing varying size of the neighborhood size k. To find the best neighborhood size of any candidate seed, they propose to compare the SN-based measure between a k-neighborhood and a 2k-neighborhood. Thresholding this relative density measure then allow to identify the cluster candidates among the candidate seeds. The main drawback of this method is that the SN-based density measure is biased towards the formation of smaller clusters over larger ones [12]. Setting the relative density threshold is also still an issue.
To overcome the cluster size bias of PatClust and most previous methods, Houle et al. [12] introduced the Relevant Set Correlation model (RSC) to measure the similarity of object sets of different sizes. RSC suggests to use Pearson correlation instead of usual intersection or cosine measure as primary SN-measure. The resulting score being fully unbiased regarding input sets sizes, this allows to compare the connectivity of clusters of different sizes and thus to optimize theoretically the neighborhood size k of any input seed object.

In all the above mentioned methods, the multi-source extension of SN clustering was not afforded. To the best of our knowledge, the more recent work dealing with several input sources of information is the one of Hamzaoui et al. [15]. They suggest to extend Houle's RSC model [12] to a single new source built from a co-occurrences based re-ranking of all available information sources. In this way, the proposed method is more an early fusion of the different sources rather than a real multi-source SN method. They show that the proposed approach succeed in fusing different heterogeneous sources but, as shown in our experiments, such method is not robust to the presence of noisy information sources (outliers). Furthermore, as it equally weights all information sources, for each candidate cluster, the fusion is far from optimal.
Using SN clustering for multi-modal clustering has been studied in several works. The more recent one is [16], authors propose a method for extracting meaningful and representative clusters that is based on a shared nearest neighbors (SNN) approach. They treat both content-based features and textual descriptions (tags) but they produce two sets of clusters (an image could be an element of a tag-based cluster and a content-based cluster). By displaing the two cluster sets with their representative form, user can browse a cluster and switch from a cluster set to another. When they combine the two clusterings (by a simply adding of the visual and textual similarity matrix), they lost the particularities of each modality and produce clusters that are less defined and hardly understandable by a user).

3 Multi-source Shared Nearest Neighbors

To generalize a shared neighbors clustering to a multi-source environment, several issues have to be solved. Whereas in single source model, each item to be clustered is associated with a single nearest neighbours list, in the multi-source environment, each item is associated with m nearest neighbors lists where m corresponds to the number of sources. In the following, we denote as F the set of available information sources and $m = |F|$. Any information source $f \in F$ is defined only by its Nearest Neighbors response function Q_f:

$$Q_f(v, K) = \{n_i\}_{i \in [1,K]}$$

where v represents any item of the whole dataset S to be clustered and $n_i \in S$ the i-th nearest neighbor of v.

3.1 From Single-source to Multi-source Shared Neighbors similarity measures

As primary Shared Neighbors similarity measure between object sets, we chose the Relevant Set Correlation (RSC) measure suggested by Houle et al. [12]. For any two sets A and B belonging to S, the RSC similarity measure is defined by:

$$R(A,B) = \frac{\mid A \cap B \mid - \frac{|A||B|}{|S|}}{\sqrt{\mid A \mid\mid B \mid (1 - \frac{|A|}{|S|})(1 - \frac{|B|}{|S|})}} \tag{3.1}$$

As already mentioned in section 2, this inter-set correlation measure improves upon previous intersection measures by removing the bias related to unbalanced set sizes [12]. From this inter-set correlation measure, Houle et al. [12] derives the following *intra-set significance* measure, for any set $A \subset S$:

$$SR(A) = \frac{1}{|A|} \sum_{v \in A} R(A, Q(v, |A|))$$

Indeed, $SR(.)$ measures the connectivity of a candidate set A as the expectation of the inter-set correlation between A and the nearest neighbors set of an item v selected uniformly at random from A:

$$SR(A) = \mathbf{E}[R(A, Q(v, |A|)) \mid v \in A]$$

We can easily extend this measure to the multi-source case, by measuring the expectation of the inter-set correlation between set A and the nearest neighbors set of an item v selected uniformly at random from A according to an information source f selected uniformly at random from F:

$$\begin{aligned} SR(A,F) &= \mathbf{E}[R(A, Q(v, |A|)) \mid v \in A \ , \ f \in F] \\ &= \frac{1}{\mid F \mid\mid A \mid} \sum_{f \in F} \sum_{v \in A} R(A, Q_f(v, |A|)) \end{aligned} \tag{3.2}$$

Unfortunately this measure has the disadvantage of a second-order bias relative to both the size of A and the number of information sources $|F|$. To show that, we can estimate the expectation and variance of $SR(A, F)$ under the hypothesis $\mathcal{H}$ that all sources are i.i.d and uniformly distributed, i.e. each source returns nearest neighbors selected uniformly at random from S. As proved in annex part at the end of this paper, under this hypothesis, we get:

$$\mathbf{E}[SR(A, F) \mid \mathcal{H}] = 0$$

which shows that SR(A,F) is not biased at the first order, thanks to the standardization introduced in the primary set correlation measure of Equation 3.1. Whereas at the second order we get

$$\mathbf{Var}[SR(A,F) \mid \mathcal{H}] = \frac{1}{|A||F|(|S|-1)}$$

which means that the standard deviation of $SR(A,F)$ under the source randomness hypothesis $\mathcal{H}$ is decreasing with both A and F. Comparing the intra-significance of sets of varying sizes and varying number of information sources is thus biased, which is problematic if we want to optimize the neighborhood size of a candidate cluster or to automatically select the optimal subset of sources, as done in the rest of the paper.
To remove this second order bias, we thus require to standardize the intra-significance $SR(.)$ thanks to $\mathbf{Var}[SR(A,F) \mid \mathcal{H}]$. Let $SI(A,F)$ be the new standardized intra-significance of set A according to information sources set F:

$$\begin{aligned} SI(A,F) &= \frac{SR(A,F) - \mathbf{E}[SR(A,F) \mid \mathcal{H}]}{\sqrt{\mathbf{Var}[SR(A,F) \mid \mathcal{H}]}} \\ &= \sqrt{(|S|-1)|A||F|}SR(A,F) \end{aligned} \tag{3.3}$$

Such standardization under a randomness hypothesis is sometimes referred as an *a contrario* method [17]. Interestingly, if we had directly standardize a simple set overlap measure $|A \cap B|$ *a contrario* to the hypothesis $\mathcal{H}$ we would have get the same result than the one obtained here by using the Set Correlation $R(A,B)$.

3.2 Source selection

Now that we have an intra-significance measure unbiased relative to the number of information sources, we can describe our approach to select the optimal subset of sources of any input set A. If we denote as $\phi \subseteq F$ an arbitrary subset of sources, then we are searching for the optimal subset $\phi_A \subseteq F$ maximizing $SI(A,\phi)$:

$$\phi_A = \underset{\phi \subseteq F}{\operatorname{argmax}} \left(SI(A,\phi)\right)$$

What seems at first glance to be a combinatorial problem can indeed be solved extremely easily by pre-sorting single-sources intra-significances; indeed, after few trivial operations, $SI(A,\phi)$ can be re-expressed as:

$$SI(A,\phi) = \frac{1}{\sqrt{|\phi|}} \sum_{f \in \phi} SI(A,\{f\}) \tag{3.4}$$

where $SI(A,\{f\})$ is the intra-significance of A according to a single source f, which can be pre-computed independently for each source $f \in F$. Now, let us decompose our maximization problem in m independent maximizations at constant $|\phi| = p$ for $p \in [1,m]$:

$$\phi_A = \underset{\phi \in \{\phi_{A,1},\ldots\phi_{A,m}\}}{\operatorname{argmax}} \left(SI(A,\phi)\right)$$

where

$$\begin{aligned} \phi_{A,p} &= \underset{\phi \subseteq F \;\; |\phi|=p}{\operatorname{argmax}} \frac{1}{\sqrt{|\phi|}} \sum_{f \in \phi} SI(A,\{f\}) \\ &= \frac{1}{\sqrt{|\phi|}} \underset{\phi \subseteq F \;\; |\phi|=p}{\operatorname{argmax}} \sum_{f \in \phi} SI(A,\{f\}) \end{aligned}$$

And if we now denote as f_i the information source $f \in F$ having the i-th highest intra-significance $SI(A, \{f\})$ we get:

$$\phi_{A,p} = \{f_i\}_{i \in [1,p]} \tag{3.5}$$

So finally, to select the optimal subset ϕ_A we just need to sort the m single-source intra-significance $SI(A, \{f\})$ in decreasing order and find the optimal number of top sources p_A:

$$p_A = \underset{p}{\operatorname{argmax}} \frac{1}{\sqrt{p}} \sum_{i}^{p} SI(A, \{f_i\})$$

Our final selected-source intra-significance measure of any arbitrary set A is finally given by:

$$SSI(A, F) = \frac{1}{\sqrt{p_A}} \sum_{i}^{p_A} SI(A, \{f_i\}) \tag{3.6}$$

4 Clustering framework

Now that we have define our new multi-source Shared Neighbors significance measures we can describe our clustering procedure, which is somehow a multi-source extension of the greedy heuristic proposed by Houle et al. [12]. It is based on two main steps, *candidate cluster selection* and *redundant clusters elimination*:

- **candidate cluster selection**: Each item $v \in S$ is considered as candidate cluster center and an optimal candidate cluster $C(v)$ needs to be computed for it. For each source $f \in F$, we first compute an optimal neighborhood $Q_f(v, k_f)$ by varying the neighborhood size k and maximizing our new selected-source intra-significance SSI (Eq. 3.6):

  $$k_f = \underset{k}{\operatorname{argmax}}\, SSI(Q_f(v, k), F)$$

  Note that during this process, an optimal source selection is performed for each iteration on the neighborhood size k and the selected subset ϕ of sources may differ from one value of k to another. Among the m optimized neighborhoods $Q_f(v, k_f)$, we finally keep as candidate cluster $C(v)$ the one with the maximum SSI score:

  $$C(v) = \underset{f}{\operatorname{argmax}}\, SSI(Q_f(v, k_f), F)$$

  We notice here that we could have further improved these candidate clusters by some *reshaping* considerations as suggested in [12] but we did not explore that track in this paper.

- **redundant clusters elimination**: We use for this step a simple heuristic based on the inter-set correlation defined in Equation 3.1. We first sort all candidate clusters $C(v)$ by decreasing order of their SSI score and then iterate on them. If an encountered cluster has an inter-set correlation higher than a user-defined threshold θ with at least one of the previously retained clusters it is eliminated. If not, it is added to the list of final clusters. We are aware that this step could be widely improved but our purpose in this paper is rather to show the contribution of the multi-source SN approach compared to the single-source one. So that the important point is that we use the same strategy for both approaches.

5 Experiments

5.1 Synthetic data

To validate our theoretical contribution, we first perform some experiments with synthetic data and information sources. We first built a synthetic set S of 5000 items clustered in $Q = 30$ categories, with category sizes varying between 20 and 200 items. Note that we do not need to build any feature vector but only sets of item identifiers. We then define three types of synthetic information sources simulating different kinds of Nearest Neighbors Responses $Q_f(v, K)$ $(v \in S)$:

- **Random** source: for any item v, $Q_f(v)$ returns items selected uniformly at random from S.
- **Perfect** source : for any item v, $Q_f(v, K)$ first returns all the items of the cluster $C(v)$ to which v belongs, with random order. Once K exceeds $|C(v)|$, it returns items selected uniformly at random from S.
- **Normal** source $N(t, r)$: a normal source is a Random source for some clusters and a noisy source for the other ones. The proportion of clusters for which it is a Random source is $(1-t)$. For the other clusters, it returns items selected uniformly at random from $C(v)$ with a probability r and items selected uniformly at random from S with probability $(1-r)$. So that, parameter t simulates the percentage of clusters for which the Normal source is *working* and parameter r simulates how well it is working for these clusters. Note that with $t = 1$ and $r = 1$, we get a Perfect source and with $t = 0$ and/or $r = 0$ we get a Random source.

We measure the effectiveness of the clustering with the two following metrics:

- **AvgPurity**: the average Purity of all returned clusters. The Purity of a cluster C is defined according to [18] by
$$Purity(C) = \frac{1}{\mid C \mid} max \mid C_h \mid$$
where C_h are the sub clusters composed by all items of C coming from the same groundtruth's category. $max \mid C_h \mid$ is thus the dominant category of the cluster.
- **F1** measure defined by
$$F1 = 2 * \frac{Prec * Rec}{Prec + Rec}$$
with
$$Prec = \frac{\#distinctclusters}{\#ofretrievedclusters}$$
and
$$Rec = \frac{\#distinctclusters}{\#ofClasses}$$
Two clusters are considered to be distinct if their dominant category differ.

Intuitively the $F1$ metric measures the ability of the clustering algorithm to retrieve the initial categories (in a data mining perspective) whereas the AvgPurity measures the semantic quality of the produced clusters.
Our first experiment consists in combining one Perfect source with $m = 5$ Random sources to validate the robustness of our source selection algorithm. As theoretically expected, our method is fully invariant to the inclusion of random sources and both F1 measure and AvgPurity are equal to 1.0.
Our second experiment is to study the influence of parameters t and r when combining several Normal information sources. For this experiment, we used systematically $n = 3$ Normal information sources and we varied the value of r and t (same values for each of the 3 independent Normal

sources). Results are provided in Table 1 and Table 2. They show that our method is robust to both kinds of noise. To strongly affect the effectiveness of our method, r and t have to be decreased to very low values, e.g 0.4 or 0.2. That means that our method is able to compensate the weak quality of very noisy independent sources by combining them effectively.
Finally, we study the impact of the number of sources on the effectiveness of our method. For this experiment we fixed r and t and we varied the number of Normal input sources from 1 to 10. We do that for two different settings, $(r = 0.6, t = 0.6)$ and $(r = 0.2, t = 0.2)$. Results are provided in Table 3. They first show that increasing the number of sources is **always** profitable which is a very consistent result for our multi-source Shared Neighbors method. For $(r = 0.6, t = 0.6)$, the errors induced by each individual source are very well compensated by only combining 5 information sources. For $(r = 0.2, t = 0.2)$ which corresponds to very weak input sources, the results remain off course quite weak even with 10 sources but the gain over a single source is still very consistent.

	t=1.0	t=0.6	t=0.4	t=0.2
r=1.0	1.0	0.96	0.88	0.42
r=0.6	1.0	0.93	0.68	0.33
r=0.4	1.0	0.82	0.62	0.28
r=0.2	0.17	0.14	0.13	0.12

Table 1: Impact of r and t noise parameters on F1 measure

	t=1.0	t=0.6	t=0.4	t=0.2
r=1.0	0.99	0.96	0.71	0.44
r=0.6	0.63	0.55	0.38	0.37
r=0.4	0.44	0.36	0.35	0.38
r=0.2	0.53	0.46	0.44	0.40

Table 2: Impact of r and t noise parameters on AvgPurity measure

		m=1	m=3	m=5	m=10
r=0.6 t=0.6	F1	0.49	0.93	1.0	1.0
	AvgPurity	0.23	0.55	0.62	0.77
r=0.2 t=0.2	F1	0.05	0.10	0.12	0.15
	AvgPurity	0.22	0.40	0.45	0.52

Table 3: Impact of the number of input information sources on F1 and AvgPurity measures.

5.2 Text-Image categorization

We performed a text-image categorization experiment based on the **Wikipedia** image dataset of ImageClef 2009 [19]. Initially, this dataset is dedicated to multimodal *retrieval* evaluatiosn but we benefit here from the provided annotations to build a text-image clustering task. Among the full $150K$ images dataset, we keep only the images that have been effectively annotated during the pooling proocedure, i.e the images that have been manually controlled as positive for at least one of the 44 query topics. The resulting dataset is composed of 1582 images categorized in 44 clusters. Each image is associated with textual information extracted from the initial Wikipedia web page (title, description, etc.). We used two information sources, one textual information source based on the TF/IDF similarity measure of PF/Tijah system [20]. One visual information source based on 5 global visual features (HSV Histogram [18], Hough histogram [18], Fourier histogram [18], edge orientation histogram[18] and probability weighted RGB histogram) and L1 metric as similarity measure.
We used the same F1 and AvgPurity metric as described before. Results are given in Table 4. The experiment demontrates the effectiveness of our approach in combining textual and visual information sources. The source selection step during the clustering process makes the results better than each single source. Clusters produced by selecting both of visual and textual sources are more semantically and visually coherent.

	Visual and Textual sources	Textual source	Visual source
F1	0.63	0.30	0.62
AvgPurity	0.17	0.51	0.35

Table 4: Clustering results on F1 and AvgPurity measures for the Sub set of Wikipedia ImageClef 2009.

5.3 Visual objects mining

We performed a visual object mining experiment based on **Caltech256** dataset. Initially, this dataset was dedicated to supervised objects classification so that clustering the 256 classes with an unsupervised method would be too difficult. We thus used this dataset in a different way to evaluate visual objects discovery in small image sets. We collected 5 subsets of the Caltech256 dataset. Each subset is construct of 10 categories and 20 random images. We used the same 5 global visual features as described in the previous experiment but in this experiment we use them as separate information sources within our multi-source Shared Neighbors framework. Table 5 demontrates the effectiveness of our approach comparing to each single source.

6 Conclusions and perspectives

Whereas Shared Nearest Neighbors methods appear to be ideally suited to **multi-modal** and heterogeneous contexts, very few works have been addressing the problem of fusing different sources of information. In this paper, we introduced a complete new **multi-source** shared neighbours framework including multi-source sets significance measures, with or without optimal source selection, and a multi-source clustering algorithm based on these new measures. We first validated the proposed theoretical contribution through original synthetic information sources experiments. They notably show that our method succeeds in increasing the clustering effectiveness simply by increasing the number of noisy and incomplete information sources. We also applied the proposed method to real multi-modal clustering benchmarks of the literature and showed that the method

		DB1	DB2	DB3	DB4	DB5
Multi-source	F1	0.38	0.66	0.56	0.27	0.57
	AvgPurity	0.57	0.70	0.36	0.67	0.59
HSV Histogram	F1	0.36	0.21	0.42	0.13	0.53
	AvgPurity	0.53	0.63	0.32	0.52	0.57
Hough Histogram	F1	0.36	0.24	0.31	0.22	0.47
	AvgPurity	0.53	0.60	0.27	0.60	0.35
Fourier Histogram	F1	0.35	0.34	0.54	0.24	0.50
	AvgPurity	0.55	0.68	0.32	0.59	0.52
Edge Orientation Histogram	F1	0.35	0.24	0.35	0.15	0.54
	AvgPurity	0.60	0.64	0.30	0.53	0.44
Prob-weighted Histogram RGB	F1	0.36	0.21	0.42	0.13	0.46
	AvgPurity	0.47	0.60	0.31	0.50	0.33

Table 5: Clustering results on F1 and AvgPurity measures for the Sub set of Caltech256.

succeeds in combining real heterogeneous and multi-modal information sources. Beyond clustering we think that the proposed approach is suitable for many other multimedia schemes, such as search results structuring and diversity enhancement, query suggestion, summarization or pattern discovery.

Acknowledgments

This work was supported under a research grant of the ANR Foundation (ANR-MDCO-2008-11/Project R2I).

References

[1] Till Quack, Bastian Leibe, and Luc Van Gool, "World-scale mining of objects and events from community photo collections," in *CIVR '08: Proceedings of the 2008 international conference on Content-based image and video retrieval*, New York, NY, USA, 2008, pp. 47–56, ACM. 4

[2] Lyndon S. Kennedy and Mor Naaman, "Generating diverse and representative image search results for landmarks," in *WWW '08: Proceeding of the 17th international conference on World Wide Web*, New York, NY, USA, 2008, pp. 297–306, ACM. 4

[3] Zheng-Jun Zha, Linjun Yang, Tao Mei, Meng Wang, and Zengfu Wang, "Visual query suggestion," in *MM '09: Proceedings of the seventeen ACM international conference on Multimedia*, 2009, pp. 15–24. 4

[4] L. Kaufman and P. J. Rousseeuw, "Finding groups in data: an introduction to cluster analysis," in *John Wiley and Sons*, New York, USA., 199. 4

[5] R. Ramakrishnan T. Zhang and M. Livny, "Birch:an efficient data clustering method for very large databases," in *Proc. ACM SIGMOD Conf. on Management of Data*, Montreal, Canada, 1996, p. 103?114. 4

[6] R. Rastogi S. Guha and K. Shim, "Cure: an efficient cluster algorithm for large databases," in *Proc. ACM SIG-MOD Conf. on Management of Data*, New York, USA, 1998, p. 73?84. 4

[7] Patrick E.A. Jarvis, R.A., "Clustering using a similarity measure based on shared nearest neighbors," *IEEE Transaction on computers*, vol. C-22(11), 1973. 4

[8] Jarvis R. Hofman, I., "Robust and efficient cluster analysis using a shared nearest neighbors approach," in *In Proc. 14th International Conference on Pattern Recognition,*, Washington D.C., 1998. 4

[9] Yu Y.Q. Zhou D.R. Meng B. Wang, H.B., "Fuzzy nearest neighbor clustering of high-dimensional data," in *International Conference on Machine Learning and Cybernetics*, 2003. 4

[10] Steinback M. Kumar V. Ertoz, L., "Finding clusters of different size, shapes and densities in noisy, high dimensional data," in *SIAM International Conference on Data Mining (SDM '03)*, 2003. 4, 5

[11] R. Rastogi S. Guha and K. Shim, "Rock: a robust clustering algorithm for categorical attributes," in *Inform. Sys. 25 (2000)*, 2000, p. 345?366. 4, 5

[12] M.E. Houle, "The relevant-set correlation model for data clustering," in *SDM*, 2008, pp. 775–786. 4, 5, 6, 8

[13] J. Sander M. Ester, H.-P. Kriegel and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. on Knowl. Discovery and Data Mining (KDD*, Portland, USA, 1996, p. 226?231. 5

[14] M.E. Houle, "Navigating massive data sets via local clustering," in *Proc. 9th ACM SIGKDD Conf. on Knowl. Disc. and Data Mining (KDD)*, 2003, p. 7547?552. 5

[15] A.Joly A.Hamzaoui and N.Boujemaa, "Multi-source rsc model for multiple search result clustering," in *Wiamis*, 2010. 5

[16] J-E.Haugeard P.A MoÃ«llic and G.Pitel, "Image clustering based on a shared nearest neighbors approach for tagged collections," in *Proceedings of the 2008 international conference on Content-based image and video retrieval, Niagara Falls, Canada*, 2008. 5

[17] Agnes Desolneux, Lionel Moisan, and Jean-Michel Morel, "A grouping principle and four applications," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 25, pp. 508–513, 2003. 7

[18] J. Z. Wang Y. Chen and R. Krovetz, "Content-based image retrieval by clustering," in *In Proceedings of the 5th ACM SIGMM international workshop on Multimedia information retrieval*, 2003, pp. 193–200. 9

[19] http://www.imageclef.org/2009/wiki, ," . 11

[20] Rode H. van Os R. Flokstra J. Hiemstra, D., "Pf/tijah: text search in an xml database system," in *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, 2006. 11

Annex 1

Assume we are given an arbitrary fixed subset $A \subset S$ and a second subset $V \subset S$ chosen uniformly at random from the items of S. Then, the random variable $|A \cap V|$ is hyper-geometrically distributed with expectation

$$\mathbf{E}\left[|A \cap V|\right] = \frac{|A||V|}{|S|}$$

and variance

$$\mathbf{Var}\left[|A \cap V|\right] = \frac{|A||V|(|S| - |A|)(|S| - |V|)}{|S|^2(|S| - 1)}$$

Consequently,

$$\mathbf{E}\left[R(A,V)\right] = \mathbf{E}\left[\frac{\mid A \cap V \mid - \frac{|A||V|}{|S|}}{\sqrt{\mid A \mid\mid V \mid (1 - \frac{|A|}{|S|})(1 - \frac{|V|}{|S|})}}\right] = 0$$

and

$$\begin{aligned}
\mathbf{Var}\left[R(A,V)\right] &= \mathbf{Var}\left[\frac{\mid A \cap V \mid - \frac{|A||V|}{|S|}}{\sqrt{\mid A \mid\mid V \mid (1 - \frac{|A|}{|S|})(1 - \frac{|V|}{|S|})}}\right] \\
&= \frac{\mathbf{Var}\left[|A \cap V|\right]}{\mid A \mid\mid V \mid (1 - \frac{|A|}{|S|})(1 - \frac{|V|}{|S|})} \\
&= \frac{1}{|S| - 1}
\end{aligned}$$

And finally, as under hypothesis $\mathcal{H}$, $Q_f(v, |A|)$ behaves as the random variable V

$$\begin{aligned}
\mathbf{E}\left[SR(A,F)\right] &= \mathbf{E}\left[\frac{1}{\mid F \mid\mid A \mid} \sum_{f \in F} \sum_{v \in A} R(A, Q_f(v, |A|))\right] \\
&= \frac{1}{\mid F \mid\mid A \mid} \sum_{f \in F} \sum_{v \in A} \mathbf{E}\left[R(A,V)\right] = 0
\end{aligned}$$

and

$$\begin{aligned}
\mathbf{Var}\left[SR(A,F)\right] &= \mathbf{Var}\left[\frac{1}{\mid F \mid\mid A \mid} \sum_{f \in F} \sum_{v \in A} R(A, Q_f(v, |A|))\right] \\
&= \frac{1}{\mid F \mid^2 \mid A \mid^2} \sum_{f \in F} \sum_{v \in A} \mathbf{Var}\left[R(A,V)\right] \\
&= \frac{\mid F \mid\mid A \mid}{\mid F \mid^2 \mid A \mid^2 (|S| - 1)} \\
&= \frac{1}{\mid F \mid\mid A \mid (|S| - 1)}
\end{aligned}$$

Unité de recherche INRIA Rocquencourt
Domaine de Voluceau - Rocquencourt - BP 105 - 78153 Le Chesnay Cedex (France)

Unité de recherche INRIA Futurs : Parc Club Orsay Université - ZAC des Vignes
4, rue Jacques Monod - 91893 ORSAY Cedex (France)
Unité de recherche INRIA Lorraine : LORIA, Technopôle de Nancy-Brabois - Campus scientifique
615, rue du Jardin Botanique - BP 101 - 54602 Villers-lès-Nancy Cedex (France)
Unité de recherche INRIA Rennes : IRISA, Campus universitaire de Beaulieu - 35042 Rennes Cedex (France)
Unité de recherche INRIA Rhône-Alpes : 655, avenue de l'Europe - 38334 Montbonnot Saint-Ismier (France)
Unité de recherche INRIA Sophia Antipolis : 2004, route des Lucioles - BP 93 - 06902 Sophia Antipolis Cedex (France)

Éditeur
INRIA - Domaine de Voluceau - Rocquencourt, BP 105 - 78153 Le Chesnay Cedex (France)
http://www.inria.fr
ISSN 0249-6399