Key Discovery for Numerical Data: Application to Oenological Practices

Danai Symeonidou[1], Isabelle Sanchez[1], Madalina Croitoru[2], Pascal Neveu[1], Nathalie Pernelle[3], Fatiha Saïs[3], Aurelie Roland-Vialaret[4], Patrice Buche[5], Aunur-Rofiq Muljarto[1], Remi Schneider[4]

[1] INRA, MISTEA Joint Research Unit, UMR729, F-34060 Montpellier, France
[2] University of Montpellier, Montpellier, France,
[3] LRI (CNRS UMR8623 & Université Paris Sud), Université Paris Saclay, Orsay, France,
[4]NYSEOS, Montpellier, France,
[5]INRA, IATE Joint Research Unit, UMR1208, F-34060 Montpellier, France

Abstract. The key discovery problem has been recently investigated for symbolical RDF data and tested on large datasets such as DBpedia and YAGO. The advantage of such methods is that they allow the automatic extraction of combinations of properties that uniquely identify every resource in a dataset (i.e., ontological rules). However, none of the existing approaches is able to treat real world numerical data. In this paper we propose a novel approach that allows to handle numerical RDF datasets for key discovery. We test the significance of our approach on the context of an oenological application and consider a wine dataset that represents the different chemical based flavourings. Discovering keys in this context contributes in the investigation of complementary flavors that allow to distinguish various wine sorts amongst themselves.

1 Introduction

Nowadays, more and more data are produced in the context of life science. Analysing and understanding such data is a task of great importance. A series of classical methods is often exploited for this analysis: [7] [8] [11]. Nevertheless, these approaches are not always explicative enough especially when data are described using many properties. To address more efficiently this problem, we propose to use a *key* discovery method. Key discovery has been initially introduced in the context of relational databases and then in the Semantic Web in order to better understand the data. A key represents a set of properties that uniquely identifies every resource in the data.

Some recent approaches [12, 6, 16, 5, 15] have been developed to discover automatically keys from RDF datasets. In this paper we use SAKey [16] that takes into account the OWA in key semantics [4]. Indeed, discovering keys in RDF datasets without taking into account RDF data specificities (such as erroneous data and duplicates) may lead to discovery of irrelevant keys or false negatives. Furthermore, certain sets of properties that are not keys but share a small number of shared values, can be useful for data linking or data cleaning. These sets of properties are particularly needed when a there are only few valid keys.

While key discovery has already been tested on large datasets with the above mentioned characteristics (such as DBpedia [2] and YAGO [3]) in SAKey [16]), many real world RDF datasets contain numerical data describing results of scientific experiments. The state of the art in key discovery in RDF datasets does not handle experimentally issued numerical data due to the following challenges that *motivate our contribution*:

- First, numerical data are too precise to enable the discovery of relevant keys. Indeed, numerical data as such hold no semantic meaning (especially when handled as Strings).
- Second, the size of the available data is in general not big enough to discover only relevant keys.

Different measures have been used in the litterature to evaluate the quality of the keys, like support and discriminability [6, 16]. However, in the context of numerical data describing scientific experiments, additional quality measures have to be defined.

In this paper we propose *a novel approach that allows to handle numerical RDF datasets for key discovery*. To deal with the numerical data, a preprocessing step is applied to convert numerical data into symbolic data. To do so, we use statistical methods to group numerical data in classes. By grouping the data into classes we exploit the closeness of values of data. The use of quantiles ensures that the values are distributed in equal-size groups. The numbers of groups were defined in order to have significant probability mass. Based on the obtained symbolic classes we apply SAKey to discover keys that are valid in this preprocessed data. Finally, two new quality measures are defined and used to evaluate the discovered keys. The first method is based on the property value distribution in the dataset and the second method is based on correlations that can be found between key properties.

In order to demonstrate the practical use of our method we test this approach on a dataset that describes wines obtained from the Pilotype project[1]. The dataset contains a set of numerical values regarding different chemical components that give the flavour of wines. In this application setting, the discovered keys can be used to discover flavour complementarity, unknown from the experts, that allow to distinguish various wine sorts amongst themselves. We have also described the dataset from a statistical point of view using principal component analysis (PCA) on the raw data (without quantile classification). This allowed us to have a more global view of the dataset and to see the way the properties (i.e., the different flavours) are correlated amongst each other. We have then validated the keys obtained with domain experts and discussed their interest with respect to the statistical analysis.

[1] http://www.qualimediterranee.fr/projets-et-produits/consulter/les-projets/theme-1-agriculture-competitive-et-durable/das2-tic-chaine-alimentaire/theme-1-developper-une-agriculture-competitive-et-durable/das-2-contribution-des-tic-a-la-chaine-alimentaire-en-amont/pilotype

The contribution of this paper is *three fold*:

- From a *methodological point of view* we extend upon the state of the art in two directions. We provide the first approach in the literature dealing with key discovery numerical data issued from experiments. We introduce novel quality evaluation criteria for such keys and discuss them in context of existing work.
- From an *application point of view* our results are extremely important as we use key discovery on (1) a domain where the interpretation of keys can only be done by domain experts and (2) for a dataset that prevents the discovery of such keys manually by the expert. This calls for both *(i)* statistical analysis to quantitatively validate the different keys found and *(ii)* a qualitative analysis to validate the interest of keys for the application.
- From an *interdisciplinary point of view* this paper successfully demonstrates the added value of putting together results issued from Computer Science, Statistics and Oenology.

This paper is organised as follows. In Section 2 we introduce the methodology proposed for the key evaluation. Then, in Section 3 we show how our method is applied in the case of a Wine dataset and we present the experimental results. In Section 4, we provide an analysis of the keys discovered and finally in Section 5 we discuss about the related works existing in this domain and we propose future works.

2 Numerical Key Discovery Methodology

A key is a set of properties that uniquely identifies every resource in a data knowledge base. In the context of Semantic Web, keys are defined for a specific class. Keys are used in several tasks such as data linking but can also be very useful to understand the data.

In this work we employ an automatic key discovery method, as the task of defining keys is very difficult for domain experts. Experts may define erroneous keys. Moreover, the more the data are described by many properties, the harder it becomes for a human expert to define all the possible keys. Please note that in experimental setting data experts may not be aware of the existence of keys. Furthermore, it has been shown in [12] that keys automatically discovered are relevant and can even be more significant than keys defined by experts.

Our numerical key discovery approach aims to discover keys in numerical experimental data. The approach is performed in several steps shown in Figure 1. In the preprocessing step, numerical data are transformed to obtain a symbolic interpretation. Then we apply SAKey to discover a set of minimal keys.

In what follows, we first present the preprocessing step, then the key discovery approach used in this work is described. Finally, several measures that can be used to evaluate the quality of the discovered keys are proposed.

2.1 Data Pre-processing

As previously explained, if we treat the data as purely numerical we get a lot of keys (since the numerical data is treated as simple Strings). Therefore, we need a way to

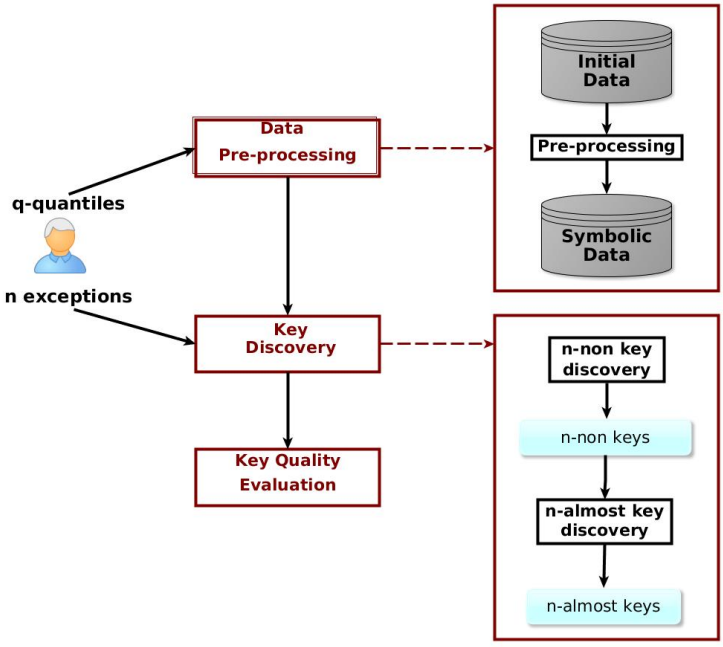

Fig. 1. Method Steps Followed In This Paper

cluster the data. Quantiles are cutpoints dividing a set of observations into equal sized groups. q-Quantiles are values that partition a finite set of values into q subsets of (nearly) equal sizes. By using quantiles, we reduce the number of values and potentially decrease the number of naive keys, so we can now take advantage of knowledge of how the properties that might work together form keys. Different strategies for computing the quantiles can be found in [11].

Intuitively, accepting only few groups may lead real keys to be not lost. On the contrary, allowing many groups can lead to the discovery of keys that are not valid in the real life. In general, more the data are grouped in few groups/many groups, more the number of properties that are involved in a key increases/decreases respectively. Therefore, choosing the appropriate size of groups can play a very significant role in the obtained results.

2.2 Key Discovery

We use SAKey [16] to discover automatically keys in the preprocessed data. SAKey is an approach that is able to discover keys in very large RDF datasets even under the presence of erroneous data or duplicates. To deal with the errors in data, SAKey discovers *almost keys*, i.e., sets of properties that are not keys due to few exceptions in the data. A set of properties is a n-almost key if there exist at most n instances that share values for this set of properties. Formally, the set of exceptions E_P corresponds to the

set of instances that share values with at least one instance, for a given set of properties P.

Definition 1. (Exception set). *Let c be a class ($c \in \mathcal{C}$) and P be a set of properties ($P \subseteq \mathcal{P}$). The exception set E_P is defined as:*

$$E_P = \{X \mid \exists Y (X \neq Y) \wedge c(X) \wedge c(Y) \wedge (\bigwedge_{p \in P} \exists U p(X,U) \wedge p(Y,U))\}$$

where X and Y are instances of the class c.

For example, in the D1 dataset showed in Figure 2 we have:

$$E_{\{d1:producer\}} = \{w1, w3\}$$

Using the exception set E_P, a n-almost keyis defined in [16] as follows:

Definition 2. (n-almost key). *Let c be a class ($c \in \mathcal{C}$), P be a set of properties ($P \subseteq \mathcal{P}$) and n an integer. P is a* n-almost key *for c if $|E_P| \leq n$.*

This means that a set of properties is considered as a n-almost key, if there exist from 1 to n exceptions in the dataset.

By definition, if a set of properties P is a n-almost key, every superset of P is also a n-almost key. Therefore, this approach discovers only minimal n-almost keys, i.e., n-almost keys that do not contain subsets of properties being n-almost keys for n fixed.

To illustrate this approach, let us introduce an example. Figure 2 contains descriptions of wines. Each wine can be described by the region it produced in, the name of the producer and finally its color.

Dataset D1:
d1:Wine(w1), $d1{:}region(w1,'' Bordeaux'')$, $d1{:}producer(w1,'' Dupont'')$, $d1{:}color(w1,'' White'')$,

d1:Wine(w2), $d1{:}region(w2,'' Bordeaux'')$, $d1{:}producer(w2,'' Baudin'')$, $d1{:}color(w2,'' Rose'')$,

d1:Wine(w3), $d1{:}region(w3,'' Languedoc'')$, $d1{:}producer(w3,'' Dupont'')$ $d1{:}color(w3,'' Red'')$,

d1:Wine(w4), $d1{:}region(w4,'' Languedoc'')$, $d1{:}producer(w4,'' Faure'')$, $d1{:}color(w4,'' Red'')$.

Fig. 2. Example Of Wine RDF Data

In this example, the property $d1{:}region$ is not a key for the class $Wine$. Indeed, there exist two wines, $w1$ and $w2$, produced in $Bordeaux$ and similarly, $w3$ and $w4$ in $Languedoc$. Considering each wine that shares the region of production with other

wines as an exception, there exist 4 exceptions for the property $d1{:}region$. Thus, the property $d1{:}region$ is a 4-almost key in this example since it contains at most 4 exceptions. Obviously, this property does not refer to a meaningful key since it contains the maximum number of exceptions (i.e., the number of exceptions is equal to the number of instances). Similarly, the property $d1{:}color$ is a 2-almost key. Regarding composite keys, the composite key $\{d1{:}region, d1{:}producer\}$ is a 0-almost key. Similarly, $\{d1{:}producer, d1{:}color\}$ is a 0-almost key.

The problem of discovering the complete set of keys automatically from the data is #P-hard [9]. Therefore, using a simplistic way to discover keys would not be appropriate. To validate if a set of properties is a n-almost key for a class c in a dataset D, a naive approach would scan all the instances of a class c to verify if at most n instances share values for these properties. Even in the cases where a class is described by few properties, the number of candidate n-almost keys can be huge. For example, let us consider a class c that is described by 60 properties. In order to discover all the n-almost keys that are composed of at most 5 properties, the number of candidate n-almost keys that should be checked would be more than 6 millions. Therefore, efficient ways to discover keys are necessary.

An efficient way to obtain n-almost keys, as already proposed in [14, 12, 16], is to discover first all the sets of properties that are not n-almost keys and use them to derive the n-almost keys. Indeed, to show that a set of properties is not a n-almost key, it is sufficient to find only (n+1) instances that share values for this set. The sets that are not n-almost keys, are defined first in SAKey as *n-non keys*. A n-non key is defined as follows:

Definition 3. (n-non key). *Let c be a class ($c \in \mathcal{C}$), P be a set of properties ($P \subseteq \mathcal{P}$) and n an integer. P is a* n-non key *for c if $|E_P| \geq n$.*

For example, the set of properties $\{d1{:}region, d1{:}color\}$ is a 2-non key (i.e., there exist at least 2 wines sharing regions and colors).

SAKey discovers maximal n-non keys, i.e., n-non keys that are not subsets of other n-non keys for a fixed n. Indeed, as shown in [14], minimal keys can be derived from maximal non keys. Once the discovery of n-non keys is completed, the approach proceeds to the next step which is the derivation of n-almost keys from the set of n-non keys. In SAKey, an efficient strategy for the derivation of n-almost keys has been introduced.

2.3 Key Quality Measures

To evaluate the quality of the discovered keys, different quality measures are proposed. Since the number of discovered keys can be numerous, strategies that allow the selection of the most significant ones are needed. Depending on the characteristics of the dataset one or more of the following quality measures can be chosen.

Key Support. To begin with, the support, a classical measure used in data mining, can be exploited. The support, denoted as *S_k*, represents the ratio of number of instances described by the set of properties involved in a key *k* to number of instances described

in the data. Intuitively, a key is significant when it is applicable to an important part of the data. Nevertheless, when all the instances are described by the same set of properties, the support cannot be exploited to compare the quality of different keys.

Key Exceptions. A second measure is the number of exceptions $|Ep|$ presented in Section 2.2 [16]. Even if keys with exceptions can be interesting, when the number of exceptions increases significantly, the quality of the discovered keys decreases.

Key Size. Keys composed of numerous properties can be difficult to evaluate. Therefore, we tend to select simple keys composed of few properties. This criteria is also taken into account in the next measure.

Key Probability. While discovering keys from the data, sets of properties that are not real keys may be discovered. More precisely, the property value distribution can be taken into account to evaluate the probability of a discovered key to have at least two instances that share values.

Given a set of properties, more it is probable to have at least two instances sharing all values for this set of properties, more this set of properties is considered significant when discovered as key [1].

For example, in a wine dataset describing 100 distinct wines, containing among others the properties $wineName$ and $YearProduction$. If we have 20 distinct wine names and 5 distinct years of production the probability that every couple of these properties is unique is very low. Therefore, if $wineName, YearProduction$ is discovered to be a key, we consider that its quality is very high.

We compute the probability using the following formula:

$$Pr_k = 1 - e^{-\frac{n(n-1)}{2p}}$$

where n is the number of instances described in the data, $p = \Pi_{i=0}^{j} m_i$ with j representing the number of properties participating in the key k and m_i is the number of distinct values of each property participating in the key.

In practice, we can set a threshold that acts as a probability lower bound that allows us to decide whether to consider a key or not.

Property Correlation and PCA. The more correlated the properties are the less informative a key is. Indeed, in such cases, its properties give intuitively the same kind of information [10]. PCA (Principal Component Analysis) is a multivariate procedure used to reduce the dimensionality of a dataset while retaining as much information as possible. PCA provides a description of the dataset by projecting the data onto new reduced dimensions that are linear combinations of the properties of the dataset. It can be used to graphically study the property correlation between properties with respect of the reduced dimensions. It also allows to visualise the individuals (the instances of the dataset and their respective values for the properties) on the reduced dimensions. However, sometimes, if the number of reduced dimensions is higher than a given threshold (still lower than the initial number of properties) the results could be difficult to interpret for a human expert.

Regarding key discovery, when the number of properties is to high, we consider that the correlation of the pairwise properties that make up a key is an indicator of how much the properties depend on each other. Please note that a correlation takes a value between -1 and 1.

3 Oenological Data Key Discovery

This section describes the practical results obtained for the method described above on a real dataset. The section is structured as follows. First, in Section 3.1 we describe the dataset used in the paper. We present in Section 3.2 the numerical experimental results: the preprocessing step of the dataset, the obtained keys and their quality measures.

3.1 Wine Dataset Description

Wine aroma data investigated in this paper have been obtained during a 4 years (2011-2014) research project called Pilotype. The aim of the project was to investigate the influence *(i)* of grape water stress on aroma precursors contents in berry and *(ii)* to study relationship between aromatic potential content in grapes and aroma profiling of corresponding wines. For this purpose, 3 different grape varieties (Chardonnay, Merlot and Grenache) from plots located in south of France (Languedoc-Roussilon region) were harvested then vinified according to three different winemaking processes. Enological, polyphenolic and aromatic analyses were performed along the winemaking process on grape, must and wine.

The data used here represent the different chemical based flavourings of wine. There are three datasets corresponding to three years of measurements: 2012, 2013, 2014. Each year measures the chemical flavouring of 63, 59 and respectively 44 wine instances. In each of the datasets used (2012, 2013, 2014), the wines are described using 19 flavourings (see column Analyzed molecules in Figure 3).

In 2013 and 2014, each grape variety was collected from at least 3 different plots in order to have triplicate conditions. Three different processes dedicated to white (Chardonnay), rosé (Grenache) and Red (Merlot) wines were used to obtain experimental wines at pilot scale (1 hL) and under standardized conditions (pilot technical cellar).

Experimental wines were analysed by classical methods (GC-MS/MS and LC-MS/MS) in order to quantify important aroma compounds as described in Figure 3. All these aroma compounds were quantified by Stable Isotope Dilution Assay (SIDA) to ensure high quality results in terms of accuracy and reproducibility.

Here, we are interested in automatically emerging any complementarity between flavours. Since, aroma compounds vary depending on the chosen year, the complementarities are studied year per year. In this context complementarity is understood in the sense that it allows for the flavours to combine in order to discriminate a fermentation method or not.

Chemical families	Concentration levels in wine	Analyzed molecules	Analysis methodology
Thiols	ppt	3MH = 3 mercaptohexanol 3MHA = 3 mercaptohexylacetate	LC-MS/MS
Esters	ppm	2PHEN= 2-phenylethanol AH= hexyl acetate AI= isoamyl acetate ABPE= phenethyl acetate DE= ethyl decanoate HE= ethyl hexanoate OE= ethyl octanoate BE= ethyl butyrate 2HPE= Ethyl lactate 3HBE= Ethyl 3-hydroxybutyrate 2MBE= Ethyl 2-methylbutyrate 2MPE= Ethyl isobutyrate 2HICE= Ethyl leucate	GC-MS/MS
C13-noriprenoïds	Ppb	BDAM= beta-damascenone BION= beta-ionone	GC-MS/MS
PDMS	Ppb	Dimethylsulfide potential= S-methylmethione + others compounds	GC-MS/MS
GSH	ppm	Glutathione	LC-MS/MS

Fig. 3. Wine Flavourings In Pilotype Dataset

3.2 Experiments

The data quantify the flavour potential of different wine varieties (expressed as instances in the RDF file). We took measurements for several different flavours thus we have a lot of flavour properties. We are interested to see if there are flavours that either go well together or are complementary. This means that when certain flavours are formed, the other flavours are not and vice-versa. Our approach will be able to extract such information from the numerical data. If a set of properties is a key that means that the flavours they represent are complementary.

Data Pre-processing. The data pre-processing step has been performed using *R* [13], a programming language and software environment for statistical computing and graphics. In this experiment, we grouped the data in quantiles using the Gumbel strategy, considering the modal position [11]. This strategy is commonly used in the R language. We grouped the datasets using 5-quantiles, 10-quantiles and finally 12-quantiles. The size of quantiles has been chosen taking into account the size of instances in the datasets. Choosing a big number of groups, may lead to not significant probability mass. In our case, each group always contains between 8% and 20% of values for a given property.

When setting the size of quantiles to less than 5, no keys were obtained. Similarly, since the datasets are describing few instances, setting a very big size of groups leads to the discovery of insignificant keys. Therefore, empirically, we selected 3 different sizes of quantiles to investigate the effect of the size in the results. Finally, the size of quantiles affects the probability score of a key.

Key Discovery. In Table 1 we show the number of discovered keys for the three datasets corresponding to the years 2012, 2013, 2014. The keys are shown for the datasets split in 5-quantiles, 10-quantiles and finally 12-quantiles.

We have discovered many minimal keys. Thus, we have used the key size quality measure to reduce the number of keys that can be shown to the expert. More precisely, for each year and each quantile only the keys that have the minimal size have been selected (see the values in bold in Table 1). In total we showed 104 keys to the expert.

Table 1. Number Of Keys Per Year And Per Quantiles In Pilotype

Nb of properties/	2012			2013			2014		
Quantiles	3	4	5	3	4	5	3	4	5
5	0	0	0	0	**2**	72	0	0	**32**
10	0	**23**	472	**3**	305	1348	**4**	454	471
12	0	**1**	149	**24**	461	705	**13**	415	664

Table 2. Quality Measures For Expert Validated Discovered Keys In Pilotype

	Year	**Quantiles**	**Support**	**Probability**	**Size**
[3MHA, BDAM, GSH, 2MPE, 3MH]	2014	5	73%	26%	5
[3MHA, GSH, OE, 2HICE, 3MH]	2014	5	73%	26%	5
[3MHA, AI, PDMS, 2MPE, 3MH]	2014	5	100%	26%	5
[3MHA, BE, PDMS, 2MPE, 3MH]	2014	5	100%	26%	5
[BDAM, OE, PDMS, 3MH]	2012	10	100%	17%	4
[GSH, OE, PDMS, 2PHEN]	2012	10	100%	17%	4
[AI, BDAM, 2HICE, 3MH]	2012	10	100%	17%	4
[3MHA, BDAM, GSH, 2MPE]	2013	5	63%	94%	4
[3MHA, BDAM, GSH]	2013	12	63%	64%	3
[AH, BDAM, GSH]	2013	12	63%	64%	3
[BE, 2HICE, 3MH]	2013	12	100%	64%	3
[BDAM, GSH, 3MH]	2013	12	63%	63%	3
[GSH, PDMS, 3HBE]	2013	10	63%	83%	3
[BDAM, GSH, 3MH]	2013	10	63%	83%	3
[GSH, PDMS, 3HBE]	2014	10	73%	63%	3
[PDMS, 3HBE, 3MH]	2014	12	100%	44%	3
[3MHA, GSH, PDMS]	2014	12	73%	44%	3
[BE, GSH, 3MH]	2014	12	73%	44%	3

Key Quality Measures. The expert has validated 18 keys among the 104 showed keys (i.e., 17.3%). In Table 2 we show for each validated key the obtained values for support, probability and size.

A set of 11 validated keys have key probability greater than 40%. This result shows that the probability measure allows to select a significant part of the relevant keys: the

smaller the probability of a set of properties being a key, the more of interest the key can be.

In the validated keys the support varies from 63% to 100% and the key probability from 17% to 94%. The fact that the expert selected a key with not a full support shows that the support as such is not a meaningful indicator of the quality of a key. Similarly we thought that the lower the number of properties in the key, the more informative the key is. However, the expert did not always follow this postulate and the best informative keys were selected as the keys of 5 properties.

To further our study, a qualitative evaluation is also available for the first three keys - the keys of size 5. They were considered by the expert as the most informative keys as each property in the key corresponds to a class of aromas in wine (i.e., same bio-genetic origin). These aromas and their pertinence as keys are better explained in Section 4.

This experiment has shown that the probability of keys is not monotonic with respect to their size. Moreover, we noticed that the probability alone is not sufficient to select all the relevant keys: the keys that were found to be the most interesting are the ones with a relatively small (but not the smallest, nor the biggest) probability. Therefore, other quality criteria should be defined such as correlation presented in Section 4.

Please note that the probabilities can serve as a way to help the expert against his /her own confirmation biases. For instance, the first key has an extremely small probability and even if the expert invariably can find a justification, from a data analysis point of view the key is artificial.

As a conclusion, on top the quantitative evaluation, a qualitative evaluation is highly needed. One reason for this is the complexity of the domain and of the data which is not easily analysed by simple numerical measures. Therefore, in the next section we will also analyse the results from a statistical point of view and put our results in context of oenological practices with a domain expert.

4 Evaluation

In this section we present the analysis of our results from a practical perspective. We have worked in close contact with one of the wine experts involved in the project Pilotype and the main results explained here are drawn after discussion with the domain expert.

4.1 Statistical Key Analysis

The dataset was analyzed using bivariate and multivariate methods. Correlations of Pearson (linearity assumption) and Spearman (without assumption) were calculated pairwise [7]. A principal component analysis was used to get an global overview of the data [8].

We consider three of the discovered keys extracted from Table 2. The first key analysed is of size three: (BE, 2HICE, 3MH). The Pearson correlation of the pairwise properties of this key vary from 0 to 0.51. The second key of size four is (AI, BDAM, 2HICE, 3MH) varies from 0.05 to 0.42. Last, the key (3MHA, BE, PDMS, 2MPE and 3MH) varies from 0.02 and 0.49. Please note that a correlation of 0.5 is the highest correlation

between these properties. Similarly, the Spearman correlation of (BE, 2HICE, 3MH) varied from 0.05 to 0.55. This confirms our hypothesis that because the properties are not significantly correlated their combination functions well as an interesting key.

We now provide the second statistical analysis that aims to multivariate analyse the data. We consider here the 2012 dataset. The PCA on our dataset on the two first principle components (Dimension 1 and 2) can explain 62% of the variability. We can see that individuals are projected into the two first axis following the oenological modality.

The information we can extract from the PCA analysis on the first and second dimension is as follows. GSH and PDMS (present in the 6th key shown in Table 2) are badly explained by the PCA detailed in Figure 4. 2PHEN is little explained by first axis and not much explained by the second axis. In general, aromatic properties are drowned within other properties and we cannot interpret them. This allows us to show the limits of the PCA analysis and its complementarity to the analysis of the key pertinence.

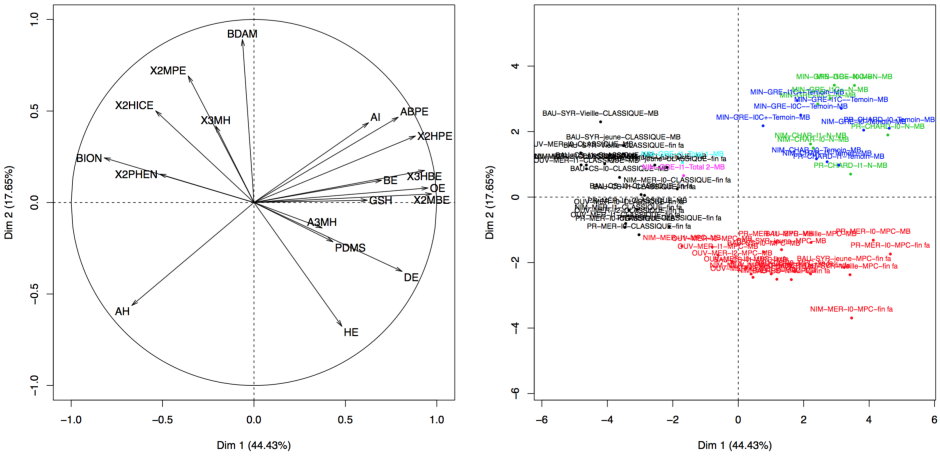

Fig. 4. PCA Of 2012 Dataset

4.2 Oenological Key Significance

Please note the description of wine flavourings in Figure 3. Here we discuss the qualitative value of the first three keys (the keys of size 5) found the most pertinent by the domain expert. These keys are composed of several aromas that are complementary to each other. Therefore they allow to better identify a wine. More precisely 3MHA and 3MH give some information on the thiol biogenesis. The higher the presence of thiol, the more the grapefruit and passion fruit like aroma is present in the wine. Such aromas are very appreciated by consumers. The PDMS gives an idea of the ageing potential of wine. The older the wine, the more DMS is released. The DMS plays an important role in the fruity taste of the wine. BE and AH are fermenting indicators and GSH gives an

idea of the level of wine oxidation. Last 2MPE gives an indication of the fermentation level of the biogenesis.

We observe that most of the obtained keys contain a class of aroma per element of the key (property). This characteristic allows the expert to define wine aromatic profiles.

5 Discussion

In a context of life science, it is important to be able to simultaneously analyse large datasets with multiple properties on large population with the aim of a comprehensive understanding of the studied system. To this end, different classical methods are available. These approaches are used for exploratory purposes in order to give more insights or descriptions into biological studies: Principal Component Analysis (global overview), Multiple Factor Analysis, Canonical Correlation Analysis or Partial Least Squares Regression (relationships of different types of properties in a non-supervised or supervised context). Sparse versions of these methods are then used in order to select relevant properties on data. This allows to explain most of the variance/covariance data structure using linear combinations of the original properties.

The great advantage of these geometrical methods is the possibility of representing the data in reduced dimensions (usually the three first components) for a global overview. Yet, to be restricted to these first components may seem too much simplified, especially since these planar visualizations often suffer of a lack of biological interpretability and require statistical background skills. The application of these kinds of methods can also be limited by the size of the dataset (particularly for CCA) and the noisy and multicolinearity characteristics of the properties (PCA, CCA, MFA). Furthermore in the case of highly dimensional case, these methods are extremely cumbersome to use.

In this paper we presented a method that allows key discovery on oenology experimental numerical data rendered symbolic by means of quantiles. We have evaluated our method practically and tested it with domain experts.

The research avenues opened by this work are numerous. First the analysis of the correlation between probabilities and the keys discovery should be taken at a larger scale and, if possible, on synthetic datasets that will allow the better analyse it. Second, we could envisage an interface with the domain expert that will facilitate the understanding of keys. This method could be generalized for association rules and dependencies allowing to discover hidden rules within the numerical data. Last, using the Pearson correlation allowed us to see that certain properties are strongly correlated. Such attributes are important to be known as they might not bring complementarity when mining for key discovery. Therefore, such information could be used in order to prune large datasets that do not scale up.

Acknowledgments

The third author acknowledges the support of ANR grants ASPIQ (ANR-12-BS02-0003), QUALINCA (ANR-12-0012) and DURDUR (ANR-13-ALID-0002). The work of the third author has been carried out part of the research delegation at INRA MISTEA Montpellier and INRA IATE CEPIA Axe 5 Montpellier.

References

1. https://en.wikipedia.org/wiki/birthday_problem.
2. http://wiki.dbpedia.org/downloads39.
3. http://www.mpi-inf.mpg.de/yago-naga/yago/downloads.html.
4. M. Atencia, M. Chein, M. Croitoru, M. L. Jerome David, N. Pernelle, F. Saïs, F. Scharffe, and D. Symeonidou. Defining key semantics for the rdf datasets: Experiments and evaluations. In *ICCS*, 2014.
5. M. Atencia, J. David, and J. Euzenat. Data interlinking through robust linkkey extraction. In *ECAI 2014 - 21st European Conference on Artificial Intelligence, 18-22 August 2014, Prague, Czech Republic - Including Prestigious Applications of Intelligent Systems (PAIS 2014)*, pages 15–20, 2014.
6. M. Atencia, J. David, and F. Scharffe. Keys and pseudo-keys detection for web datasets cleansing and interlinking. In *EKAW*, pages 144–153, 2012.
7. Chen P.Y. and Popovitch P.M. *Correlation: Parametric and Nonparametric Measures. Sage University Papers Series on Quantitative Applications in the Social Sciences*, 2002.
8. F. Husson, S. Lê and J. Pagès. *Analyse de données avec R, 2ème édition revue et augmentée*, 2016.
9. D. Gunopulos, R. Khardon, H. Mannila, S. Saluja, H. Toivonen, and R. S. Sharma. Discovering all most specific sentences. *ACM Trans. Database Syst.*, 28(2):140–174, June 2003.
10. S. Holmes. Multivariate analysis: the french way. pages 1–14, 2006.
11. R. J. Hyndman and Y. Fan. Sample quantiles in statistical packages. *The American Statistician*, 50:361–365, 1996.
12. N. Pernelle, F. Saïs, and D. Symeonidou. An automatic key discovery approach for data linking. *J. Web Sem.*, 23:16–30, 2013.
13. R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2015.
14. Y. Sismanis, P. Brown, P. J. Haas, and B. Reinwald. Gordian: efficient and scalable discovery of composite keys. In *VLDB*, pages 691–702, 2006.
15. T. Soru, E. Marx, and A.-C. Ngonga Ngomo. ROCKER – a refinement operator for key discovery. In *Proceedings of the 24th International Conference on World Wide Web, WWW 2015*, 2015.
16. D. Symeonidou, V. Armant, N. Pernelle, and F. Saïs. SAKey: Scalable Almost Key discovery in RDF data. In *Proceedings of the 13th International Semantic Web Conference (ISWC2014)*, ISWC '14. Springer Verlag, 2014.