The Computational Wine Wheel 2.0 and the TriMax Triclustering in Wineinformatics

Bernard Chen(✉), Christopher Rhodes, Alexander Yu, and Valentin Velchev

Department of Computer Science,
University of Central Arkansas, 201 Donaghey Ave, Conway, AR 72034, USA
bchen@uca.edu

Abstract. Even with the current state of technology, data growth is increasing so fast that without proper storage and analytical techniques, it is challenging to process and analyze large datasets. This applies to knowledge bases from all fields and all kinds of data. In Wineinformatics, various kind of data related to wine, including physicochemical laboratory data and wine reviews, are analyzed by data science related researches. In the previous work, we proposed the Computational Wine Wheel, derived from 2011's top 100 wine, to automatically process and extract key attributes from human-language-format wine expert reviews. In this work, past 10 year's top 100 wines are collected and formed a 1000 excellent wines dataset to further improve the Computational Wine Wheel. The extraction process led to the creation of what we call a Computational Wine Wheel 2.0, which is a wine attribute dictionary consisting of 985 categorized and normalized wine attributes. After the Computational Wine Wheel 2.0 is formed, we experiment it on a region- and grape type- specific dataset to seek new types of information in Wineinformatics. A novel TriMax Triclustering algorithm specifically used for the dataset processed by the Computational Wine Wheel is proposed and applied to discover three dimensional clusters (Wine × Attributes × Vintage) in wine. We found that the TriMax Triclustering algorithm produced promising and cohesive results that can be used in various aspects of the wine industry, such as defined palate grouping and wine searching.

Keywords: Wineinformatics · The computational wine wheel · Biclustering · Trimax Triclustering

1 Introduction

There is an intrinsic notion that the computational power of today is essentially limitless, especially when we realize that today's cell phones have more computational power than all of NASA had when it landed two astronauts to the moon in 1969 [1]. We can only imagine what future computational power will be like given said power is supposed to double every eighteen months according to Moore's Law. Even with contemporary capabilities though, it would seem that we could process anything imaginable. However, with more computational power comes the ability to actually

© Springer International Publishing Switzerland 2016
P. Perner (Ed.): ICDM 2016, LNAI 9728, pp. 223–238, 2016.
DOI: 10.1007/978-3-319-41561-1_17

generate new and vastly-growing data every single day. So much data in fact that it is estimated we will have generated 40 zettabytes of data by the year 2020 [2]. With ever-growing sizes in raw data, we have problems not only parsing the data itself, but pulling out meaningful information from it as well. At its core, Data Science is the study that incorporates varying techniques and theories from distinct fields, such as Data Mining, Scientific Methods, Math and Statistics, Visualization, natural language processing, and the Domain Knowledge, to discover useful information from domain-related data. Among all fields in the study of data science, the domain knowledge is the starting point as well as the ending point since all data science researchers need to start with the domain problem, and end with useful information within the domain.

Wine was considered as a luxury in old days; however, it is more and more popular and enjoyed by a wide variety of people today. U.S. consumers bought 29.1 million hectoliters of wine in 2013, a rise of 0.5 % on 2012, while French consumption fell nearly 7 % to 28.1 million hectolitres [3]. Because of the popularity, the demand for luxury and high quality wines produced in great years is high despite their high price; for example, Chateau Petrus 2009 costs $45,600 per case before tax and sold out before the release date. Fortunately, for consumers' point of view, tens of thousands of wines are produced per year and the quality of the wine is not reflecting merely based on the price.

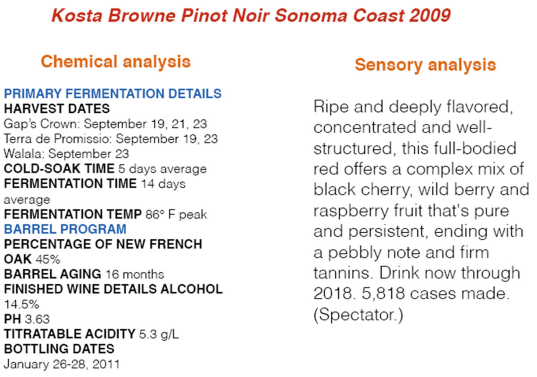

Fig. 1. The review of the Kosta Browne Pinot Noir Sonoma Coast 2009 (scores 95 pts) on both chemical and sensory analysis

The quality of the wine is usually assured by the wine certification, which is generally assessed by physicochemical and sensory tests [4]. Figure 1 provides an example for a wine review by both perspectives. Physicochemical laboratory tests [5, 6] routinely used to characterize wine include determination of density, alcohol or pH values, while sensory tests rely mainly on human experts [6]. Most of the existing data mining/data science researches related to wine [6–8] focus on the physicochemical laboratory tests data, which is stored in the UCI Machine Learning Repository. However, in wine economics point of view, sensory analysis is much more interesting to consumers and industrial perspective than chemical analysis since they describe aesthetics, pleasure,

complexity, color, appearance, odor, aroma, bouquet, tartness, and the interactions with the senses of these characteristics [9] of the wine.

The source of the dataset is always an important factor to the success of a research. Chemical analysis data comes from the lab and costs about $1000 per wine; Sensory analysis produces by prestigious experts who generate consistent wine sensory reviews. In United States, several popular wine magazines provide widely accepted sensory reviews toward wines produced every year, such as Wine Spectator [14], Wine Advocate [15] and Decanter [16] etc. All of those wine magazines review thousands of wines through the 100-point scale and testing notes, which is in the human language format as showed in Fig. 1. Currently, the Wine Spectator database holds more than 300,000 wine reviews. Unquestionably, from these large amount of data, it is interesting to discover meaningful information from those sensory testing notes for answering the questions such as "What makes wine achieve a 90 + rating and considered as a outstanding wine?", "What are the common characteristics shared by 90 + Napa Cabernet sauvignon?", "What characteristics differ between wines from Bordeaux, France and Napa, United States?"

In our previous work [10] published in December 2014, the term "Wineinformatics" was proposed to apply data science techniques and natural language processing on professional wine reviews. A Computational Wine Wheel based on 2011's top 100 wines is proposed to automatically extract wine attributes from professional reviews. The work has been cited in different wine related researches, including mobile app development [11], financial prediction [12] and accessing wine quality [13] research area. In this paper, we would like to redefine the Wineinformatics as a study that incorporate data science in any wine related dataset, including physicochemical laboratory data and wine reviews.

This paper will present the Computational Wine Wheel 2.0 for extracting key attributes from wine reviews like the Fig. 1 example. We will detail the formation of a Computational Wine Wheel 2.0, which will serve as a basis for future, automated extraction of attributes from wine reviews. Given the dictionary and a couple datasets of wine reviews, we explore varying clustering techniques in an attempt to show that it is possible to group similar wines together using only the sensory attributes. A novel tri-cluster in wine (Wine × Attributes × Vintage) is also proposed in this paper. We believe our examination and subsequent evaluation of wine sensory information can advance Wineinformatics researches.

2 The Computational Wine Wheel 2.0

2.1 Wine Sensory Reviews

The wine testing process can be very delicate as a wine is examined not only for its tasting quality, but for physical appearance and physiochemical properties as well. A taster will usually evaluate the appearance of the wine, how it smells in the glass before tasting, the different sensations once tasted, and finally how the wine finishes with its aftertaste. The taster will be looking for how complex the wine is, how much potential it has for aging for drinkability, and if there are any faults present. The experience required can be expansive

as any given wine needs to be carefully assessed within comparable wine standards according to its price, region, varietal, and style. Also, if known, the actual wine production techniques will allow the taster to examine further characteristics. To show an example of what might result from a professional tasting, below is an example wine tasting review for Wine Spectator's number one wine of 2014.

> ***Dow's Vintage Port 2011***
> *Powerful, refined and luscious, with a surplus of dark plum, kirsch and cassis flavors that are unctuous and long. Shows plenty of grip, presenting a long, full finish, filled with Asian spice and raspberry tart accents. Rich and chocolaty. One for the ages. Best from 2030 through 2060.*

Similar reviews are provided by various prestigious wine magazines [14–16]. Among those, we chose Wine Spectator as our primary data source to start aggregating our wine reviews because of their strong on-line wine review search database and consistent wine reviews. These reviews are mostly comprised of specific tasting notes and observations while avoiding superfluous anecdotes and non-related information. They review more than 15,000 wines per year and all tastings are conducted in private, under controlled conditions. The magazine has been in production since 1976. The company has their reviews available for subscribers directly to their website. Wines are always tasted blind, which means bottles are bagged and coded. Reviewers are told only the general type of wine and vintage. Price is also not taken into account. Their reviews are straight and to the point. For each reviewed wine, a rating within a 100-points scale is given to reflect how highly their reviewers regard each wine relative to other wines in its category and potential quality. The score summarizes a wine's overall quality, while the testing note describes the wine's style and character. The overall rating reflects the following information recommended by Wine Spectator about the wine [14]: 95 ~ 100 Classic; 90 ~ 94 Outstanding; 85 ~ 89 Very good; 80 ~ 84 Good; 75 ~ 79 Mediocre; 50 ~ 74 Not recommended.

In the review example listed above, the attributes are neatly stated without much confusion to what constitutes a proper wine tasting note. For example, to manually process the review, all the terms that are bold will be extracted and considered characteristics of the wine:

> ***Dow's Vintage Port 2011***
> ***Powerful, refined*** *and* ***luscious,*** *with a surplus of* ***dark plum, kirsch*** *and* ***cassis*** *flavors that are* ***unctuous*** *and* ***long.*** *Shows plenty of* ***grip,*** *presenting a* ***long, full finish,*** *filled with* ***Asian spice*** *and* ***raspberry tart*** *accents.* ***Rich*** *and* ***chocolaty.*** *One for the ages. Best from 2030 through 2060.*

We have bolded key attributes, and these attributes range from actual savory properties, such as "chocolate" and "Asian spice", to subjective properties, such as "powerful" and "refined." One of our major research goals in this paper is to extract as many key attributes as possible from these professional wine reviews automatically.

2.2 The Computational Wine Wheel 2.0

The tasting notes given in a review are very important as they describe the heart and soul of a wine. Even without knowing the producer or varietal, a well-described review can adequately sway a potential consumer into a purchase. Our idea is to build a Savory Wine Dictionary where common, yet important attributes can be stored and referenced as needed. Luckily, this idea was introduced by a sensory chemist and professor named Nobel [17]. She created the Wine Aroma Wheel which is composed of twelve categories of overall wine aromas someone might experience when tasting a wine. Without being overly specific there are times when certain distinct flavor attributes are not unique enough to encapsulate all flavors. An example of this would be the FRUITY → (TREE) FRUIT → APPLE attribute. As we will show later with our expansion attributes, things like APPLE and GREEN APPLE are unique enough to warrant a distinction in the (TREE) FRUIT subcategory. Besides, Nobel's wine aroma wheel describes only actual savory attributes; the adjectives and wine body attributes are not included. If we map Novel's wine aroma wheel to the previous example, the processed wine savory review will be:

> ***Dow's Vintage Port 2011***
> *Powerful, refined and luscious, with a surplus of* ***dark plum, kirsch*** *and* ***cassis*** *flavors that are unctuous and long. Shows plenty of grip, presenting a long, full finish, filled with* ***Asian spice*** *and* ***raspberry tart*** *accents. Rich and* ***chocolaty.*** *One for the ages. Best from 2030 through 2060.*

By expanding the wine aroma wheel, we developed the first version of the Computational Wine Wheel based on Wine Spectator 2011's Top 100 wine and presented it in 2014 International Workshop on Domain Driven Data Mining [10]. The work has been referenced in several related area including mobile app development, financial prediction and accessing wine quality [11–13]. In order to automatically capture as many important characteristics as possible of wine reviews, we advanced the Computational Wine Wheel into the next level in this paper. To achieve the goal, we build our new Computational Wine Wheel based on TEN times more wine reviews from Wine Spectator's Top 100 Wines of 2003 to 2013 for a much comprehensive dictionary. The extraction process for these reviews was purely manual as we handpicked key attributes as well as noted secondary information about the wine. The idea of the Computational Wine Wheel is to memorize the results human labor works and compose a domain specific dictionary for human language processing. In total we gathered the following information: name, vintage, review, varietal, regional information, and price. However, it is worth noting that for our processing purposes the review is the single most important piece of information for a wine. For the review and attributes themselves, there were a few types of attributes we are concerned with. Besides actual biological flavor attributes, we also tried to include anything corresponding to a wine's physical structure, including things like acidity, body, structure, weight, tannins, and finish. These are properties of wine that a taster will physically taste or feel, such as how acidic the wine tastes or how well the wine coats the tongue. Lastly, we also decided to keep generic, subjective terminology that may or may not be the same between two different tasters. For example, one taster may find a wine "vivid" and "beautiful" while another taster may make no mention.

Showing the previous example review again, we want to highlight how we would extract the review's key attributes into the three mentioned categories: savory, body, and descriptive.

Dow's Vintage Port 2011

Powerful!, refined! and luscious!, with a surplus of dark plum, kirsch* and cassis* flavors that are unctuous! and long!. Shows plenty of grip+, presenting a long+, full finish+, filled with Asian spice* and raspberry tart* accents. Rich! and chocolaty*. One for the ages. Best from 2030 through 2060.*

For this review, **red(*)** words indicate specific flavors and aromas that could possibly be found on Nobel's wine aroma wheel. **Orange(+)** words indicate traits corresponding to the physical wine itself like its body and finish. That is, how the wine feels physically to a taster. Lastly, **blue(!)** words indicate subjective adjectives used by the taste to describe the overall wine. Should a word or phrase not exist in the original wine aroma wheel, we would add it. Also, if a word or phrase does not fit into any previous categories or subcategories, we would create one for it.

After we process all 1000 wine reviews, we found out there was some contextual overlap between different reviews. That is, there would be two different reviews using slightly different words to express the same tasting notes. A simple example would be one review using the word "distinctive" and another review saying a wine was "very distinct." The human thought process would naturally assume these two differences are the same thing, but computationally, we might miss the connection. For this reason, we added a FOURTH level to the wine aroma wheel that we like to call a **normalized attribute** name. This portion of the wheel would represent a base, or normalized, word to encompass a variety of word usages. This is extremely important not only for differences in word tense or suffixes, but especially the verbiage used when describing biological elements like fruits and their descriptions. A good example of this would be "blueberry", "blueberry fig", and "blueberry jam." Even though all three are components of the same fruit, the taste and consistency of each item convey different connotations and perceptions. All of these normalized processes require domain expert to make judgments. Luckily, our team has a domain expert to assist us.

The Computational Wine Wheel proposed in this paper ended up with 14 distinct categories and a total of 34 distinct subcategories, which is the same with the first version of the Computational Wine Wheel. From all wines mined, we found a total of 1881 specific wine attributes, and of those attributes we were able to finalize 985 distinct normalized attributes. Table 1 provides a detail comparison between the original Computational Wine Wheel and the new one. We also identify the plurals problem in this paper, "BLUEBERRY" and "BLUBERRIES" should be treated as the same specific term as well as normalized attributes. This required program to identify the plurals for specific terms. The full Computational Wine Wheel is available under: https://dl.dropboxusercontent.com/u/13607467/CWW2.0_nonplural.txt.

Table 1. Comparison of the old and new computational wine wheel

	The computational wine wheel 2.0	The computational wine wheel
Data source	Past 10 years top 100 wines	2011 top 100 wines
Categories	14	14
Subcategories	34	34
Specific terms	1881	635
Normalized attributes	985	444
Plurals	Yes	No

2.3 How to Use the Computational Wine Wheel

In order to clarify the usage of the computational wine wheel, we provide an example in this subsection. Table 2 gives a simplified computational wine wheel, which contains only 6 specific attributes.

Table 2. Simplified computational wine wheel

Category	Subcategory	Specific Attribute	Normalize Attribute
FRUITY	BERRY	RASPBERRY	RASPBERRY
FRUITY	BERRY	RASPBERRY TART	RASPBERRY TART
FRUITY	TROPICAL FRUIT	DARK PLUM	PLUM
FRUITY	TROPICAL FRUIT	PLUM	PLUM
OVERALL	FLAVOR/DESCRIPTORS	RICH	RICH
OVERALL	FLAVOR/DESCRIPTORS	RICH AROMAS	RICH

Here is the process of how we apply the simplified computational wine wheel on the ***Dow's Vintage Port 2011*** wine review: The very first step is to use the words in the Specific Attribute column, which is the 3rd column in Table 2, to scan the review starting with the longest number of combination word. Since the longest number of combination word in the example is 2, we start with Raspberry Tart, followed by Dark Plum, and Rich Aromas. For every word scan, if we had a hit, the wine will have a positive attribute in the corresponding Normalized Attribute and remove the word from the review. Therefore, after the scan of "Raspberry Tart", we got a hit from our review; the wine will have a positive value of "Raspberry Tart" attribute. After the scan of "Dark Plum", we got a hit from our review; the wine will have a positive value of "Plum" attribute. After the scan of "Rich Aromas", we got a miss from our review; the wine will have a negative value of "Rich" attribute.

Once the highest number of combination word is processed, we scan the next number of combination word; in this example, we scan the single word Specific attribute with the same logic. Table 3 represents the Dow's Vintage Port 2011's wine

attributes in binary format after the process mentioned above. Please note that the RASPERRY attribute is still negative since we delete the word "RASBPERRY TART" from the review during the first scan. The readers may notice that many important attributes in the example are NOT included, such as ASIAN SPICE, CHOCHLATY… etc. It is because the computational wine wheel is the simplified version. The more SPECIFIC and NORMALIZED attributes included in the computational wine wheel, the more attributes can be picked up from the wine reviews to produce more accurate results. This is also the main reason that we proposed the Computational Wine Wheel to provide higher quality of natural language processing on wine reviews.

Table 3. Attributess of the processed wine example

RASPBERRY	RASPBERRY TART	PLUM	RICH
0	1	1	1

2.4 New Napa Cabernet Sauvignon Dataset Automatically Processed by the Computational Wine Wheel

The quality of the wine is based on various influences; however, two of the most well know and probably most important factors are soil and weather. Soil (or terror) reflects the characteristics of the region and depend on the composition of the soil. Weather controls the quality of the grape production and it changes every year. To recognize and study both factors in Wineinformatics, we collect region specific wine savory reviews over years and compose a dataset processed by the proposed Computational Wine Wheel automatically.

The new dataset encompasses 50 Cabernet Sauvignon wines from the Napa Valley region in California, which is one of the most famous wine regions in United States. For every wine in this set, we retrieved its review for every year from 2006 to 2010. In other words, 250 (50 wines $\times$ 5 years) wine reviews are processed by our Computational Wine Wheel. In this way, we control the soil factor and discuss the wines with different weather condition over different years (vintages). Although the dataset may look small, it is caused by our strict criteria: grape type (Cabernet Sauvignon), wine production region (Napa Valley) and years (a wine must have complete wine reviews throughout all years in research). Some wines share the same producer, but each wine has a distinct designation and is technically considered as a different wine production. For this dataset it is best to imagine it as a three dimensional cube of reviews, where the height, width, and depth are the wine name, attributes, and vintage, respectively. This dataset is special as there was nothing manual about attribute extraction. We used the computational wine wheel and scripted the output of only matched attributes. The result of this was 50 wines with 259 attributes across 5 years. The main purpose of this dataset is to discover similar wines over years under specific conditional control criteria. More detail is discussed in the following section.

3 Triclustering

Clustering is generally considered an unsupervised learning and analysis tool. The Wineinformatics information retrieved by clustering can be beneficial to various roles involved in wine. Based on wine consumers' favorite wine, wines in the same cluster can be recommended. Wine retailers can decide to purchase similar wines as the strength of the store or choose representative wines in different cluster to increase the diversity. Wine makers will be able to find wines with similar characteristic within the winery or the wine region to review their wine making process.

The classical clustering algorithms, such as hierarchical clustering and k-means clustering, are usually very good places to start when attempting to explore data. Hierarchical clustering has been successfully applied and visualized in Wineinformatics researches [10]. However, they are flawed in a sense as both algorithms are attempting to detect patterns in observations across all given attributes of a dataset. Sometimes it might be more important to find patterns that consist of a subset of attributes.

In Wineinformatics, the dataset processed through the Computational Wine Wheel is clearly a sparse binary dataset. As the result, dimension selection plays an important role in this research. A bicluster is equivalent to a biclique in a corresponding bipartite graph. This essentially means that all of a bicluster's rows, or observations, are all connected to every column, or attribute, presented in the bicluster. The idea of a bicluster should be explored as it presents the opportunity to find subspaces in our data where subsections of columns define a cluster instead of all attributes contained from that cluster's wines.

The BiMax BiClustering algorithm was a reference method developed by Prelic et al. for baseline comparison of biclustering algorithms in general [19]. The process is fairly simple in that it searches for biclusters that consist entirely of 1 s in a binary matrix. This is perfect for datasets generated with the computational wine wheel in mind because a wine fits the binary requisite; a wine either has an attribute or it does not. With this in mind, our goal is to use the BiMax algorithm to find all inclusion-maximal biclusters of wines and attributes. This means a bicluster cannot be fully contained within another bicluster.

Just as with biclustering, triclustering is becoming a popular method to explore gene expression microarray data with additional "time" dimension. Instead of working with two dimensional matrices, triclustering focuses on finding behavioral patterns between row and columns along a time series. Existed works in the gene expression field for triclustering include Zhao and Zaki [20] and Bhar et al. [21]. However, none of their dataset is considered as sparse binary dataset; hence, the developed algorithm cannot be applied in Wineinformatics directly.

This paper proposes a novel TriMax TriClustering reference algorithm specifically for sparse binary dataset. Trimax Triclustering should be considered a reference algorithm in that it attempts to cluster on the most basic level and makes no assumptions of differing values in the data. That means it expects all values to either be zero or non-zero, so completely binary in nature. For our specific dataset, all data

values are either 1 or 0. We consider a tricluster (W, A, T) to correspond to a subset of wines $W \subseteq \{1...n\}$ that jointly share a subset of wine attributes $A \subseteq \{1...m\}$ across a subset of time $T \subseteq \{1...o\}$. The tuple (W, A, T) $(W, A, T) \in 2^{(1...n)} \times 2^{(1...m)} \times 2^{(1...o)}$ is considered inclusion maximal if and only if it meets the following two criteria:

(1) $\forall i \in W, j \in A, k \in T : e_{ijk} = 1$
(2) $\exists (W', A', T')$ with (a) meets criteria (1) and (b)

$W \subseteq W' \wedge A \subseteq A' \wedge T \subseteq T' \wedge (W', A', T') \neq (W, A, T)$ Criteria (1) states that given a possible tricluster, every possible value must be a 1 across all rows, columns, and time slices. Criteria (2) is the inclusion-maximal stipulation that says a tricluster A is considered inclusion-maximal as long as there does not exist another tricluster B in which the grouping of wines, attributes, and time slices of A are a subset of B. If a tricluster A is found, there also cannot be a tricluster B, such that A = B. Now that we have defined a tricluster, we can discussed the algorithm to find them. However, there should be two points noted before we discuss the algorithm. (1) Our proposed algorithm uses the BiMax algorithm, so a good understanding of the algorithm, as we discussed in the previous section, is necessary to proceed. (2) We believe our program is able to find all triclusters, but unlike BiMax which knows at runtime whichs biclusters to ignore thanks to its column callstack, TriMax has to filter out duplicate or subset triclusters after finding all possible triclusters. We will examine an example dataset that shows how duplicates arise, but first we will run through the algorithm itself.

The pseudocode for our proposed TriMax TriClustering algorithm is shown in Fig. 2. As a base concept, we want to take biclusters found in each time slice and see if they can extend across any and all other time slices. To accomplish this, we start with our dataset D and process each of D's t time slices iteratively. For a given bicluster b that is found in a given time slice t, we form a new dataset D', which consists of the rows and columns of b, along every time slices of the input data. That new dataset is then recursively processed using the same methodology until the resulting dataset D' consists only of values of 1. Naively, we can consider a completely 1-valued dataset as a tricluster if it passes the minimum row, column, and time slice amounts set in $mWAT$. Since our process does not have any callstacks like the BiMax algorithm, TriMax will natively introduce duplicate triclusters or triclusters that are subsets, or non-maximal. To combat part of this problem, we introduce a visited array vT, which is populated with the index of a time slice once that time slice's recursive processing has finished. This allows any tricluster found to be ignored if it includes a time slice within vT at any recursive level. If this occurs, ideally it means that the tricluster has already been found previously. However, this only attempts to filter out triclusters between given time slices. It does not work on duplicate or non-maximal subsets formed from partially overlapping biclusters originating from the same time slice. Figure 3 shows an example of duplication issues caused by overlapping biclusters in a given time slice T1.

In Fig. 3, we can say we would process time slice T1 first by expanding the three biclusters found within it: {W1, W2, W3, W4} × {A1}, {W2, W3} × {A1, A2, A3}, and {W2, W3, W4} × {A1, A2}. As shown by the blue squares, all three biclusters share the following subset of rows and columns: {W2, W3} × {A1}. By expanding all

TriMax TriClustering Reference Algorithm

$ID = D = \{\{w_i\}_{i=1}^{a}\}_{i=1}^{t}$
$TriList = \{\emptyset\}$
TriMaxTriClust$(D, mWAT, vT)$:
1 **for any** $t \in D_T$ **where** $t \in vT$ **and all** $D_{W,A,t} = 1$
2 | **return**
3 **for** $t \in D$ **do**:
4 | **if**$(t \in vT$ **or** $(all\ values\ in\ t = 1\ or\ 0))$:
5 | | **if**$(all\ values\ in\ t = 0)$:
6 | | | $append(vT, t)$
7 | | **continue to next time slice**
8 | $B =$ ***BiMaxBiClust***$(t, mWAT)$
9 | **for** $b \in B$ **do**:
10 | | $newD' = \{b\}_{t=1}^{ID_T}$
11 | | ***TriMaxTriClust***$(D', mWAT, vT)$
12 | $append(vT, t)$
13 **if**$(len(D_W, D_A, D_T) \geq \{mWAT\})$:
14 | $append(TriList, D_{W,A,(D_T - vT)})$: **unless** $D_T - vT = \{\emptyset\}$
15 **return**
16
17 $\forall\ t \in TriList$: *remove duplicates/subsets*

Fig. 2. Proposed TriMax TriClustering reference algorithm pseudocode

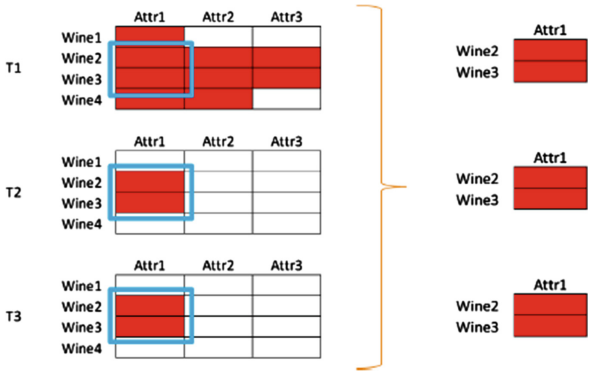

Fig. 3. Tricluster found from multiple intra-timeslice biclusters

three, the same tricluster, as presented on the right of Fig. 4, will be found three times, and thusly will have to be filtered down to one instance afterwards. Even with the slight timing inefficiency here in the post processing, we believe this method will still find all maximal triclusters between all-time slices given in a three dimensional data set. The next section will detail the results when applying triclustering to our multi-vintage 50-wine dataset.

4 Results

4.1 Biclustering 50 Wines

Due to the fact that the biclustering algorithm works on two dimensional data, we arbitrary choose 2010 vintage as the input dataset. For the 50 wines in the 2010 vintage, we implemented and applied the BiMax biclustering. The overall bicluster summarization is described in Fig. 4. This figure represents the total number of maximal biclusters found for the dataset. The table values represent the total number of biclusters that share a specific number of wines (vertical axis) versus a specific number of savory attributes (horizontal axis). For example, in this vintage there are 16 clusters that have exactly four wines and four attributes. In the table, there are darkened, rectangular borders that are meant to be a visual reference to show all biclusters where the minimum number of rows equals the minimum number of columns. For a dataset with 50 wines and 259 possible attributes, this may seem like a low combination, but it makes sense as the number of total possible attributes in any given wine is fairly small.

We then attempt to explore those biclusters that fall into the category of at least 4 wines and 4 attributes. In total, there were 17 (16 + 1) total biclusters that fell into this group, which represents some of the most robust biclusters from this vintage, region, and varietal. Table 4 shows an example of a bicluster that has four wines and share four common attributes. In this example, distinctive flavors that a taster might be accustomed to when sampling a Cabernet Sauvignon. This bicluster was also described as RICH and DENSE as well. Unlike hierarchical clustering, which would present groups of wines using all attributes among them, biclustering allows us to show many different, but smaller, groupings of the same wines across varying attribute patterns. This would give potential for consumers to select small flavor profiles and expect higher

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	21	TOTAL
1	0	0	0	0	0	1	1	1	7	3	8	6	3	6	4	3	2	2	1	1	49
2	1	31	64	108	50	18	10	0	1	0	0	0	0	0	0	0	0	0	0	0	283
3	7	50	74	38	9	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	179
4	9	39	33	16	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	97
5	4	26	23	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	54
6	2	16	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	23
7	7	6	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	20
8	3	5	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
9	1	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	6
10	1	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	9
11	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
12	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	4
14	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
15	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	2
16	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
18	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
19	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
21	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
22	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
27	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
28	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1
TOTAL	48	193	207	163	59	20	11	1	8	3	8	6	3	6	4	3	2	2	1	1	749

Fig. 4. Summarization of biclusters of 50 Wines in 2010 vintage

quality results since the biclusters might have filtered out unneeded attributes. The business value of this specific cluster might be "the customer who enjoy one of the four wines, may also enjoy other three wines" and "the customer who like the combination of the four wine attributes can be recommended those four wines".

Table 4. An example of bicluster that has four wines and share four common attributes.

Wine	Shared attributes
Chappellet signature	Black licorice
Beringer private reserve	Rich
Araujo eisele vineyard	Dark berry
Cavus stags leap district	Dense

We can also look for interesting wine and attribute combinations in those clusters that have either low number of wines and high number of attributes, or those with high number of wines and low number of attributes. The former suggests a smaller subset of wines stayed consistent across a majority of attributes across vintages, while the later suggests a larger subset of wines that might share a small pool of distinctive attributes.

4.2 Triclustering 250 Wines

To test out the TriMax Triclustering algorithm, we used the full dataset described in Sect. 2.4 and we found 23,225 possible triclusters. Since we knew a large percentage of these would actually be duplicates or non-maximal, we performed the pairwise subset comparison and pulled out a total of 7,296 superset triclusters. Of all the triclusters found, 6,357 of them only exist in a single time slice, which means these clusters are discoverable simply through Biclustering algorithm. We found 735 triclusters that spanned 2 time slices (Fig. 5A). We found 166 triclusters that spanned 3 time slices (Fig. 5B), and 31 triclusters that spanned 4 time slices (Fig. 5C). Lastly, we found 7 triclusters that spanned all 5 time slices (Fig. 5D). In Fig. 5, the darkened, rectangular borders that are meant to be a visual reference to show all clusters where the minimum number of rows equals the minimum number of columns in each time slice.

To understand the meaning of Fig. 5, let us look into an example. The "1" circled in Fig. 5D indicates there is one tricluster where 8 wines share a single attribute across all five vintages. Through the example in Table 5, the tricluster lets us know that all eight of these wines are considered GREAT for five years in a row. We can also see that four of the wines share the same producer, so it is probable that any other wine produced by BOND would also probably be considered great. One may argue that the attribute GREAT maybe not a significant word. However, if a wine can be reviewed with GREAT in their review sentences, the wine usually scores pretty well. Furthermore, this tricluster example also shows a good reason of why we categorize our attributes in

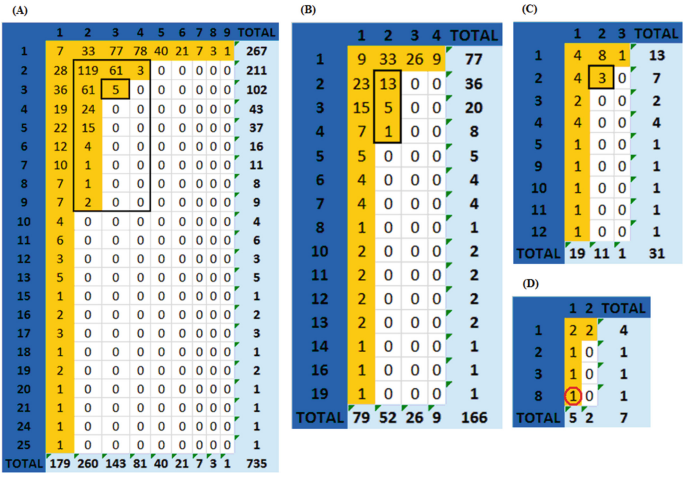

Fig. 5. Summarization of Triclusters of 250 wines in (A) two years (B) three years (C) four years (D) five years time slices

the computational wine wheel into flavor and aroma, body and finish, overall adjectives (red, orange, blue in Sect. 2.2 respectively). If the user is NOT interested in certain category of attributes, they may turn off the attribute detection during the attribute retrieve process.

Table 5. Tricluster example where eight wines share one common attribute across five different vintages

Wine	Attributes	Vintage
BARNETT spring mountain district rattlesnake hill Beringer private reserve, bond melbury, bond quella, bond st. eden, bond vecina, diamond creek gravelly meadow, diamond creek volcanic hill	Great	2010 2009 2008 2007 2006

Table 6. Tricluster example where eight wines share one common attribute across five different vintages

Wine	Attributes	Vintage
Casa piena Dancing hares	Blackberry Great	2006, 2007, 2008, 2009

Table 6 shows another tricluster that has two wines containing two attributes across four consecutive vintages. It is one of three triclusters appeared in the middle rectangular in Fig. 5C. While both wines and attributes are still fairly small in number, this just further provides opportunity for specialized searching and classification. Since the dataset used

for this paper contained only a specific varietal from NAPA, we were able to get highly defined cluster results. We believe that the triclusters discovered from a variety of types and sources should produce interesting results and it will be worth exploring those datasets in the future.

5 Conclusion

Data Science is a successful study that incorporates varying techniques and theories from distinct fields. This paper propose the Computational Wine Wheel 2.0 composed by 1000 outstanding wine reviews to support the new data science application field named Wineinformatics, which is a novel data science application area proposed by this paper. We also make the Computational Wine Wheel 2.0 publically available. A new Napa Valley Cabernet Sauvignon across five vintages dataset is automatically generated via the Computational Wine Wheel 2.0. BiMax Biclustering algorithm is applied to the dataset to find similar wines with precise amount of wine attributes. We also develop a new TriMax Triclustering algorithm specifically designed for Wineinformatics. It is the first time that tricluster is applied to Wineinformatics field to form Wine $\times$ Attribute $\times$ Vintage clusters. The novel idea is able to discover similar wines under different weather condition in different years. We believe this paper helps define the role of Wineinformatics. Many other similar fields such as coffee, whisky and chocolate with professional reviews can also follow the concept and methods in this paper to construct new data science application fields.

References

1. Kaku, M.: Physics of the Future: How Science Will Shape Human Density and Our Daily Lives by the Year 2100. Doubleday, New York (2011)
2. Gantz, J., Reinsel, D.: The digital universe in 2020: big data, bigger digital shadows, and biggest growth in the far east. IDC December 2012
3. USA becomes world biggest wine market as French drinkers cut down, 13 May 2014. http://www.reuters.com/article/2014/05/13/us-wine-usa-france-idUSKBN0DT0YO20140513. Accessed March 2015
4. Ebeler, S.: Linking flavor chemistry to sensory analysis of wine. In: Flavor Chemistry - Thirty Years of Progress pp. 409–422. Kluwer Academic Publishers (1999)
5. Chemical analysis of grapes and wine: techniques and concepts. Patrick Iland Wine Promotions, Campbelltown, Australia (2004)
6. Cortez, P., Cerdeira, A., Almeida, F., Matos, T., Reis, J.: Modeling wine preferences by data mining from physicochemical properties. Decis. Support Syst. **47**(4), 547–553 (2009)
7. Ishibuchi, H., Nakashima, T., Nii, M.: Classification and Modeling with Linguistic Information Granules: Advanced Approaches to Linguistic Data Mining. Springer, Heidelberg (2005)
8. Ishibuchi, H., Yamamoto, T.: Rule weight specification in fuzzy rule-based classification systems. IEEE Trans. Fuzzy Syst. **13**(4), 428–435 (2005)
9. Olkin, I., Lou, Y., Stokes, L., Cao, J.: Analyses of wine-tasting data: a tutorial. J. Wine Econ. **10**(01), 4–30 (2015)

10. Chen, B., Rhodes, C., Crawford, A., Hambuchen, L.: Wineinformatics: applying data mining on wine sensory reviews processed by the computational wine wheel. In: 2014 IEEE International Conference on Data Mining Workshop (ICDMW), pp. 142–149. IEEE, December 2014
11. Kiselev, A., Kuznetsov, A.: Developing a mobile application for wine amateurs (2015)
12. Zhou, Y.: Research on the applications of data mining in financial prediction (2015)
13. Lee, S., Park, J., Kang, K.: Assessing wine quality using a decision tree. In: 2015 IEEE International Symposium on Systems Engineering (ISSE), pp. 176–178. IEEE, September 2015
14. Wine Spectator Magazine. http://www.winespectator.com/. Accessed March 2015
15. rRobertParker. http://www.erobertparker.com/info/wineadvocate.asp. Accessed March 2015
16. Decanter.com. http://www.decanter.com/wine. Accessed March 2015
17. Nobel, A.C.: N.d., Wine Aroma Wheel. http://winearomawheel.com/. Accessed 29 March 2015
18. Wine Spectator: N.p., n.d., Top 100 List. http://top100.winespectator.com/lists/. Accessed 29 March 2015
19. Prelic, A., Bleuler, S., Zimmermann, P., Wille, A., Bühlmann, P., Gruissem, W., Hennig, L., Thiele, L., Zitzler, E.: A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics **22**(9), 1122–1129 (2006)
20. Zhao, L., Zaki, M.J.: TRICLUSTER: an effective algorithm for mining coherent clusters in 3D microarray data. In: SIGMOND 2005 (2005)
21. Bhar, A., Haubrock, M., Mukhopadhyay, A., Wingender, E.: Application of a novel Triclustering method (delta-TRIMAX) to mine 3D gene expression data of breast cancer cells. In: GCB 2013 (2013)