VOL 2 (2018) NO 3 - 2

e-ISSN : 2549-9904
ISSN : 2549-9610

INTERNATIONAL JOURNAL ON INFORMATICS VISUALIZATION

Patient's Cancer Categorization Based on Jaccard and Venn Techniques

Nurul Aswa Omar [#], Riswan Efendi [*], Norfaradilla Wahid [#], Afiq Luqman Mohd Yasin [**],

[#] *Department of Web Technology, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia*
[*] *Department of Mathematics, UIN Sultan Syarif Kasim Riau, Pekanbaru, 28293, Indonesia*
[**] *Department of Software Engineering, Faculty of Computer Science and Information Technology, Universiti Tun Hussein Onn Malaysia, Malaysia*
E-mail: nurulaswa@uthm.edu.my, riswan.efendi@uin-suska.ac.id, faradilla@uthm.edu.my, afiq_luqman95@yahoo.com.my

***Abstract*— Under some circumstances, hospitals and clinics need to know the preliminary patient's disease categorization, such as H1N1, Dengue, HIV/AIDS and others. However, categorizing patients will be useful in the future if the hospitals properly recorded their patient criteria and crucial details such as patient's name, symptoms, temperature and others. There are many benefits by grouping the patients together such as the hospitals will know how to provide an appropriate medicine to the patient, which patient will get the highest priority and should be quarantined. It is very important to know how to group the patients by using their symptoms detected. This study will focus on patients with cancer problems using Jaccard index and Venn diagram techniques. Their data will be collected from a hospital and the patients are selected randomly. Patients with lungs, brain and breast cancer will be selected in this study.**

***Keywords*—Cancer, Jaccard index, Similarity, Venn diagram, Categorization.**

I. INTRODUCTION

Many techniques have been implemented to classify the type and level (stage) of patient's cancer, such as, pathway analysis [7], network clustering [7], machine learning [8], and others. However, it is not an easy task to classify the cancer patients because the similarity symptoms between each cancer.

This paper will show simple technique to organize the data, using similarity value, namely, Jaccard Index. Besides that, we also use the Venn's Diagram to show how this technique works. By using Jaccard index and Venn's Diagram techniques, we can determine the relations between patients and their symptoms using set operations, such as, intersection, join and disjoint. This making prediction of each patient disease will be much easier next time and benefit to the hospitals have the required data by looking at the past diagram / pattern.

Having lots of patients will make it harder for the hospitals to remember and predict the past record. Recording and analysing patient data is easier when categorized the data using Jaccard index and Venn diagram techniques.

II. FUNDAMENTAL CONCEPT OF JACCARD INDEX AND VENN DIAGRAM

A. Jaccard Index

Paul Jaccard was a professor of botany and plant physiology at the ETH Zurich. He studied at the University of Lausanne and ETH Zurich (PhD 1894). He continued studies in Paris with Gaston Bonnier. He developed the Jaccard index of similarity (he called it coefficient de communauté) and published it in 1901 [1]. He also introduced the use of the species-to-genus ratio (he called it generic coefficient) in biogeography [2]. In the 1920s, Paul Jaccard engaged himself in a dispute with the Finnish botanist and phytogeographer Alvar Palmgren over the interpretation of species-to-genus ratio, as evidence of competitive exclusion (as held by Jaccard) or attributable to random sampling (as held by Palmgren) [3].

The Jaccard Index, also known as Intersection over Union and the Jaccard similarity coefficient (originally coined coefficient de communauté by Paul Jaccard [1]), is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard coefficient measures similarity between finite sample sets, and is defined as the

size of the intersection divided by the size of the union of the sample sets:

$$J(A,B)=\frac{|A\cap B|}{|A\cup B|}=\frac{|A\cap B|}{|A|+|B|-|A\cap B|} \quad (1)$$

If A and B are both empty, we define J(A,B) = 1 and $0\le J(A,B)\le 1$.

The Jaccard distance, which measures dissimilarity between sample sets, is complementary to the Jaccard coefficient and is obtained by subtracting the Jaccard coefficient from 1, or, equivalently, by dividing the difference of the sizes of the union and the intersection of two sets by the size of the union:

$$d_J(A,B)=1-J(A,B)=\frac{|A\cup B|-|A\cap B|}{|A\cup B|} \quad (2)$$

B. Venn Diagram

John Venn, FRS [4-5], FSA [6], (4 August 1834 – 4 April 1923) was an English logician and philosopher noted for introducing the venn diagram, used in the fields of set theory, probability, logic, statistics, and computer science.

John Venn was born on 4 August 1834 in Kingston upon Hull, Yorkshire to Martha Sykes and Rev. Henry Venn, who was the rector of the parish of Drypool [7].

A Venn diagram (also called primary diagram, set diagram or logic diagram) is a diagram that shows all possible logical relations between a finite collections of different sets. These diagrams depict elements as points in the plane, and sets as regions inside closed curves. A Venn diagram consists of multiple overlapping closed curves, usually circles, each representing a set. The points inside a curve labelled *S* represent elements of the set *S*, while points outside the boundary represent elements not in the set *S*. Thus, for example, the set of all elements that are members of both sets *S* and *T*, $S \cap T$, is represented visually by the area of overlap of the regions *S* and *T*. In Venn diagrams the curves are overlapped in every possible way, showing all possible relations between the sets. They are thus a special case of Euler diagrams, which do not necessarily show all relations. Venn diagrams were conceived around 1880 by John Venn. They are used to teach elementary set theory, as well as illustrate simple set relations in probability, logic, statistics, linguistics and computer science.

III. Descriptive Statistics for Patients Cancer

Some patient data are used to construct an information table. From the table we construct, we can convert to graph or diagram by using Jaccard or Venn. Data that has been collected is consist of the total patient's, age, gender, cancer type, and stage. We can have more data but for this documentation we will focus on this main data first. We will use 20 patients randomly in this study.

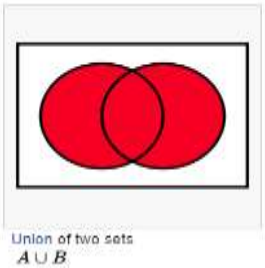

Fig. 1 A sample Union of two sets

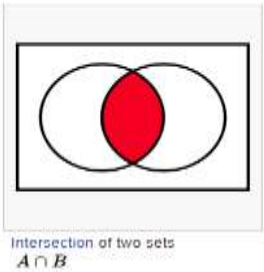

Fig. 2 A sample Intersection of two sets

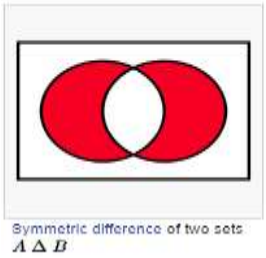

Fig. 3 A sample difference of two sets

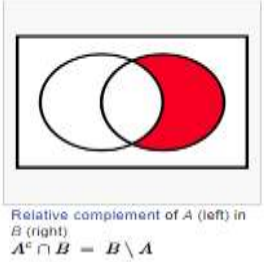

Fig. 4 A sample complement of sets

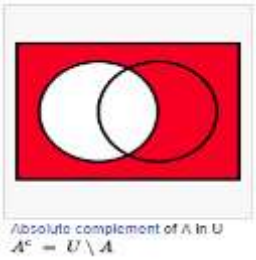

Fig. 5 A sample absolute complement of two sets

TABLE I
PATIENTS WITH CANCER INFORMATION

Patients	Age	Gender	Cancer	Stage
Patient 1	45	Male	Lung	1
Patient 2	55	Female	Brain	1
Patient 3	53	Male	Lung	2
Patient 4	76	Male	Lung	2
Patient 5	64	Male	Lung	1
Patient 6	62	Female	Breast	1
Patient 7	57	Female	Breast	2
Patient 8	49	Male	Brain	1
Patient 9	54	Female	Lung	1
Patient 10	56	Female	Breast	2
Patient 11	66	Female	Breast	2
Patient 12	61	Male	Lung	2
Patient 13	60	Female	Brain	1
Patient 14	52	Male	Lung	1
Patient 15	62	Female	Brain	2
Patient 16	71	Male	Brain	1
Patient 17	70	Male	Lung	2
Patient 18	63	Male	Lung	2
Patient 19	78	Female	Breast	1
Patient 20	68	Male	Lung	2

Table 1 shows the data of patients with cancer collected for this case study. They have different cancer type and stages making it suitable to be converted as graph or diagram to see its pattern.

By extracting the data from the Table 1, from 20 patients we can get 11 patients that are males and the other 9 is females, which is 5 patients with breast cancer, 10 patients with lung cancer, and 5 patients with breast cancer. We have the data now let's convert it to a graph (Fig. 6) to make it easy to look at.

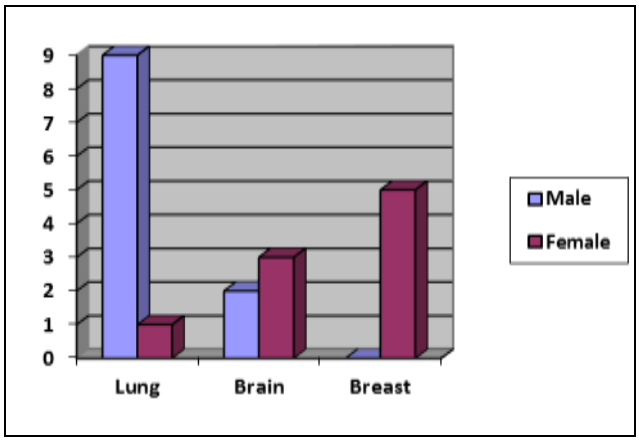

Fig. 6 The bar graph shows the number of patients with their cancer type respectively.

Fig. 6 shows that the lung cancer of male is highest among brain and breast cancers if compared with female. While, female has the highest breast cancer if compared with male. Fig. 7 shows the number of male and female who have cancers with stages 1 and 2 are same. Fig. 8 shows the bar chart patients cancer based on age and its gender. In majority, number of female was higher than male for group age 51-60 years old. While, male patient with cancer was greater than female for group age 61-70 years old significantly in this figure.

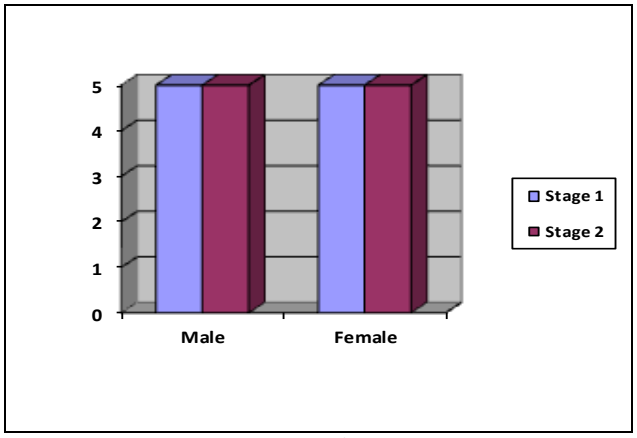

Fig. 7 The bar graph shows the number of patients with their cancer stage respectively.

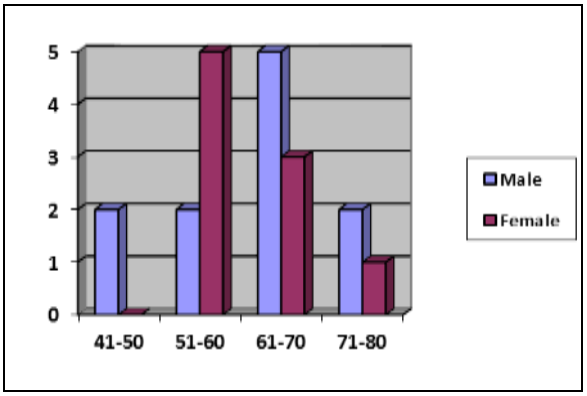

Fig. 8 The bar graph shows the number of patients in their age range.

IV. Jaccard Index and Venn Diagram for Cancer Categorization

A. Categorization Using Jaccard Index

Case study 1: Stage 1 and lung cancer

1) Step 1: Based on Table 1, define set *A* is a set of patients with stage 1 cancer and *B* is a set Patients with lung cancer as follows:

A = {Patient 1, Patient 16, …, Patient 2, Patient 19},

B = {Patient 17, Patient 1, …, Patient 9, Patient 12}.

2) Step 2: Based on Step 1, determine intersection between sets *A* and *B* as follows:

$A \cap B$ = {Patient 1, Patient 5, Patient 9}, and

$A \cup B$ = {Patient 17, Patient 1,…, Patient 19, Patient 12}.

By using Eq. (1), we have

$$\frac{|A \cap B|}{|A \cup B|} = \frac{3}{4} = 0.214 .$$

So the similarity of set *A* and set *B* is 0.214 or 21.4%. We can create many relations that we need between sets by assigning data to a specific set. Similarity between sets is important because similarity measure or similarity function is a real-valued function that quantifies the similarity between two objects.

The importance of similarity between 2 sets using Jaccard index is we can find the percentage of the 2 sets, as in the above example, we knows that only 21.4% of patients have lung cancer in stage 1. Doctors need the percentage to do their research or prepare the report. Let's move on to the next example.

Case study 2: All patients and cancer stage 1

1) Step 1: Based on Table 1, define set *A* is a set of all patients and *B* is a set patients with cancer stage 1 as follows:

A = {Patient 3, Patient 17,…, Patient 4, Patient 12},

B = {Patient 1, Patient 13,…, Patient 5, Patient 9}.

2) Step 2: Based on Step 1, determine the intersection between sets A and B as follows:

$A \cap B$ = {Patient 1, Patient 13, …, Patient 5, Patient 9},and

$A \cup B$ = {Patient 3, Patient 17,…, Patient 4, Patient 12}

$$\frac{|A \cap B|}{|A \cup B|} = \frac{10}{20} = 0.5 \, or \, 50\% .$$

Therefore, we can conclude that 50% or half of the patients are having their stage 1 cancer. We can see how important it is to use Jaccard index to get the information that we need accurately and in number form.

B. Categorization Using Venn Diagram

Venn diagram focuses on visual instead of mathematical operations like Jaccard Index making it highly preferable among researchers because of its easiness of use and classifications.

Let's say we take all of the patients and the number of patients with their cancer type as the first example.

U : All patients
A : No. of patients with lungs cancer
B : No. of patients with brain cancer
C : No. of patients with breast cancer

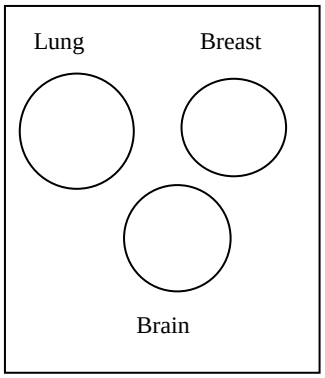

Fig. 9 The diagram shows all patients with their cancer respectively

We have the rough diagram now but what is the relationship? Well we will construct the relationship using the Fig. 9 we had just now to get the necessary output that we need. But first, we need to know the information that we need, in the next example we will find how many male and female patients having their stage 1 or 2 cancers.

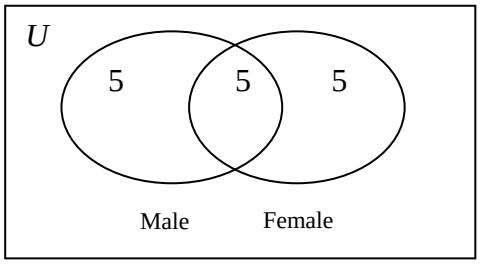

Fig. 10 The diagram shows number of male and female patients having their cancer stage. The intersection is stage 2 patients.

From Fig. 10, we can conclude that 5 male and 5 female patients are in stage 1 cancer, while the other 5 male and 5 female patients are in stage 2 cancer or vice versa. It is much easier to see the data in diagram form. Let's draw another Venn diagram using this situation, number of male and female patients that are having brain cancer.

Let's draw another Venn diagram using this situation, number of male and female patients that are having brain cancer.

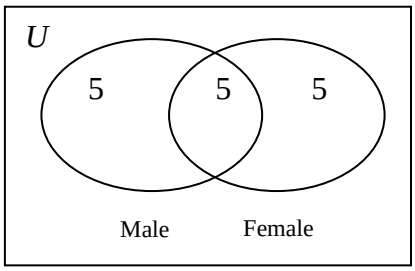

Fig. 11 The diagram shows number of male and female patients having their cancer stage. The intersection is stage 2 patients.

Fig. 11 is identical with Fig 10, there are 5 patients from male and female with brain cancer while the other 5 males and 5 females are having other type of cancer. But Venn Diagram is not suitable for all of condition such as in this case study compared to Jaccard Index. You need more information to make Venn Diagram worth to solve this case study.

V. CONCLUSION

Patient records are collected in groups such as the type of cancer, gender and stage of the disease. This is important for hospital records especially when it involves statistics. The results of this record can generate reports of many different types of bar graphs or venn diagrams using stored records.

Based on this study, Jaccard technique will give you the answer in number especially in fractions and percentage while Venn will give you a descriptive through diagram. Both of them are useful when it comes to write a report but choosing the best approach is crucial because the reader will understand more if we provide a good material. Venn diagram and Jaccard technique have their own benefit and drawback intrinsically Venn diagrams are pleasing to our eyes and easy to spot while Jaccard will provide you a strong data and reliable numbers.

REFERENCE

[1] Jaccard, P. "Distribution de la flore alpine dans le bassin des Dranses et dans quelques régions voisines" in *Bulletin de la Société Vaudoise des Sciences Naturelles,* 1901, pp. 241-272.

[2] Jaccard, P. "Étude comparative de la distribution florale dans une portion des Alpes et des Jura" in *Bulletin de la Société Vaudoise des Sciences Naturelles*, 1901, pp. 547-579.

[3] Järvinen, Olli "Species-To-Genus Ratios in Biogeography: A Historical Note" in *Journal of Biogeography* 1982, pp. 363-370.

[4] *Anonymous,* "Venn biography". *School of Mathematics and Statistics University of St Andrews, Scotland,* 2003, Retrieved 2014-08-03.

[5] *Anonymous,* "John Venn – Mathematician Biography, Facts and Pictures". *Famous-mathematicians.com.* Retrieved 2014-08-03.

[6] *Pickles, John D. "Venn, John Archibald (1883–1958)" in Oxford Dictionary of National Biography*. Oxford University Press. doi:10.1093/ref:odnb/40972.

[7] Fung, DC, Lo A, Jankova, L, Clarket, SJ, Molloy, M, Rubertson, GR, Wilkins, MR. Classification of cancer patients using pathway analysis and network clustering. *Methods Mol Biol*., 2011, 781, pp. 311-336.

[8] Chip, M. L, Victor, HVB, Herman, BF. Application of unsupervised analysis techniques to lung cancer patient data. *PLoS ONE,* 2017, 12(9), pp. 1-18.