Unilateral Jaccard Similarity Coefficient

Julio Santisteban
Universidad Católica San Pablo
Campus Campiña Paisajista s/n Quinta Vivanco,
Barrio de San Lázaro
Arequipa, Peru
jsantisteban@ucsp.edu.pe

Javier L. Tejada Carcamo
Universidad Católica San Pablo
Campus Campiña Paisajista s/n Quinta Vivanco,
Barrio de San Lázaro
Arequipa, Peru
jtejadac@ucsp.edu.pe

ABSTRACT

Similarity measures are essential to solve many pattern recognition problems such as classification, clustering, and retrieval problems. Various similarity measures are categorized in both syntactic and semantic relationships. In this paper we present a novel similarity, Unilateral Jaccard Similarity Coefficient (uJaccard), which doesn't only take into consideration the space among two points but also the semantics among them.

Categories and Subject Descriptors

E.1 [**Data Structures**]: Graphs and networks; G.2.2 [**Graph Theory**]: Graph algorithms

General Terms

Theory

Keywords

Jaccard, distance, similarity

1. INTRODUCTION

Since Euclid to today many similarity measures have been developed to consider many scenarios in different areas, particularly in the last century. Similarity measures are used to compare different kind of data which is fundamentally important for pattern classification, clustering, and information retrieval problems [3]. Similarity relations have generally been dominated by geometric models in which objects are represented by points in a Euclidean space [12]. Similarity is defined as "Having the same or nearly the same characteristics" [4], while the metric distance is defined as "The property created by the space between two objects or points". All metric distance functions must satisfy three basic axioms: minimality and equal self-similarity, symmetry, and triangle inequality.

$$d(i,i) = d(j,j) \leq d(i,j) \tag{1}$$

$$d(i,j) = d(j,i) \tag{2}$$

$$d(i,j) + d(j,k) \geq d(i,k) \tag{3}$$

Here for objects i, j and k, where d() is the distance between objects i and j. Bridge [1] argues that there exists empirical evidence of violations against each of the three axioms. Yet, there also exists geometric models of similarity which take asymmetry into account [10]. Nosofsky points out that a number of well-known models for asymmetric proximity data are closely related to the additive similarity and bias model [5]. Tversky [13] has proposed a different model in order to overcome the metric assumption of geometric models. One of the strengths of contrast models is its capability to explain asymmetric similarity judgments. Tversky's asymmetry may often be characterized in terms of stimulus bias and determined by the relative prominence of the stimuli.

$$sim(a,b) = \frac{|A \cap B|}{|A \cap B| + \alpha|A - B| + \beta|B - A|}, \quad \alpha, \beta \geq 0 \tag{4}$$

Here A and B represent feature sets for the objects a and b respectively; the term in the numerator is a function of the set of shared features, a measure of similarity, and the last two terms in the denominator measure dissimilarity: α and β are real-number weights; when α != β. Jimenez et al. [6], Weeds and Weir [14] and Lee [7] also propose an asymmetric similarity measure based on Tversky's work. However all proposals include a stimulus bias, asymmetric similarity judgments, which Tversky refers to as human judgment. Today, similarity measure is deeply embedded into many of the algorithms used for graph classification, clustering and other tasks. Those techniques are leaving aside the semantic of each vertex and it's relation among other vertices and edges.

In a direct graph, the similarity from U to Z is not the same as the distance from Z to U, this due to the intrinsic features of a direct graph. The similarities are different because the channels are dissimilar. According to Shannon's information theory we could argue that each vertex is a source of energy with an average entropy which is shared among it's channels, and while that information flow among the vertex's channels, we need to be consider it in the similarity. A similarity does not fit all tasks or cases.

In Natural Langue Processing, where the similarity between two words is not symmetric sim(word a,word b) != sim(word b,word a). WordNet [4] presents 28 different types of relations; those relations have direction but are not symmetrical, they are not even synonyms because each synonym word has

SIGIR Workshop on Graph Search and Beyond '15 Santiago, Chile
Published on CEUR-WS: http://ceur-ws.org/Vol-1393/.

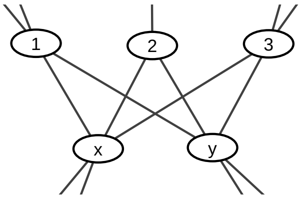


Figure 1: Structural Equivalence.

a particular semantic, meaning and usage, but are similar. Hence if two words have symmetric distance or similarity, those two words are the same. Paradigmatic is an intrinsic feature in language, It lets the utterer exchange words with other words, words with similar semantics [11]. In this paper we focus on paradigmatic analysis to support our unilateral Jaccard Similarity coefficient (uJaccard).

The rest of the paper is organized as follows. In section 2 we will show the unilateral Jaccard Similarity coefficient (uJaccard). In section 3 we will consider some cases; finally in section 4 we conclude this work.

2. PARADIGMATIC SIMILARITY DEFINITION

2.1 Basics Of Paradigmatic Structures

Paradigmatic analysis is a process that identifies entities which are not related directly but are related by their properties, relatedness among other entities and interchangeability [2]. In language the reason why we tend to use morphologically unrelated forms in comparative oppositions is to emphasize the semantics, this is done by substitution and transposition of words with a similar signifier. Similarity is not defined by a syntactic set of rules but rather by the use of the language. In some cases this use is not grammatically or syntactically correct but it is commonly used. We defined the signifier as being the degree of relation among entities of the same group, where not all members of the group have the same degree of relatedness. This is due to the fact that a member of a group might belong to more than one group.

2.2 Extended Paradigmatic

Two vertices in a graph are structurally equivalent if they share many of the same network neighbours. Figure 1 depicts a structural equivalence between two vertices y and x who have the same neighbours. Regular equivalence is more subtle, two regularly equivalent vertices do not necessarily share the same neighbours, but they do have neighbours who are themselves similar [8] [15]. We will use structural equivalence as the bases of uJaccard.

2.2.1 *Unilateral Jaccard Similarity*

To calculate a paradigmatic similarity we start with a question, is the similarity coefficient from vertex Va to Vc the same to the similarity coefficient from vertex Vc to Va ?. If we argue that both similarity coefficients are the same, we are arguing that the edges from the vertices Va and Vc are the same, and it is clear that that is not usually the case. Thus both vertices have different sets of edges. One problem with Tversky [13] similarity is the estimation for α and β which are stimulus bias, generally a human factor. Similarly, other similarities which are based on Tversky idea, have the same problem. On the other hand we propose a measure that does not include this bias. We propose a modified version of Jaccard Similarity coefficient (1), unilateral Jaccard Similarity coefficient (uJaccard) (2)(3), used to identify the similarity coefficient of Va to Vc With respect to vertex Va, and to also identify the similarity coefficient of Vc to Va With respect to vertex Vc.

$$Jaccard(V_a, V_c) = \frac{|a \cap c|}{|a \cup c|} \tag{5}$$

$$uJaccard(V_a, V_c) = \frac{|a \cap c|}{|edges(a)|} \tag{6}$$

$$uJaccard(V_c, V_a) = \frac{|c \cap a|}{|edges(c)|} \tag{7}$$

Here Va and Vc are the number of edges in vertex a and c, likewise the edges(Vc) are the number of edges in vertex c. if uJaccard is close to 0, it means that they are not similar at all. The objective of using uJaccard is to identify how similar a vertex is to other vertices in relation to itself. uJaccard could be calculated among two connected vertices, uJaccard could also be calculated among vertices that are not connected directly, but which are connected by in-between vertices. The number of in-between vertices could be from 1 to n, we do not recommend a deep comparison since the semantics of the vertex loosest its meaning. Hence max(n)=3, it is suggested for NLP. For the calculations we do not consider the number of in-between vertices since we focus on the information flow and not the information transformation carried out on the intermediate vertices.

3. EXPERIMENTAL EVALUATION

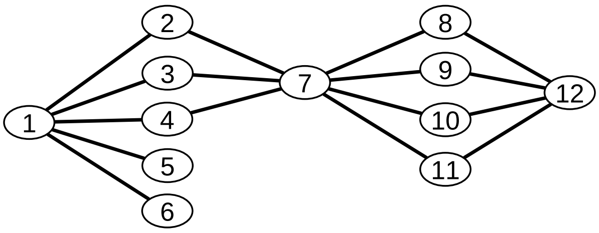


Figure 2: Toy graph.

3.1 Toy Testing

Using similarity uJaccard (6),(7) we can build a paradigmatic approach to group vertices. Figure 2 shows a toy graph with 12 vertices and 16 edges, following the paradigmatic analysis, we can determine that vertex 12 and 7 belong to group P because they have the same number of edges to a same set of vertices. Vertex 1 also belongs to group P because vertex 1 has 3 of the 5 edges, the same as vertex 7, the degree of membership of vertex 1 is lower than vertices 7 and 12 because vertex 1 has other edges that are not shared by vertices 7 or 12. In the same manner we can determine that vertex 8, 9, 10 and 11 belong to group Q because they have an equal number of edges to the same set of vertices. Similarly vertices 2, 3 and 4 belong to group R, and vertices 5 and 6 belong to group O. In this example we can easily identify the paradigmatic approach, where two or more vertices belong to the same group if they have the same or similar neighbours, but the neighbours in turn belong to another group.

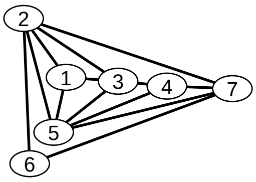
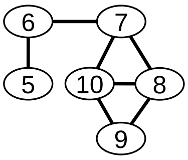
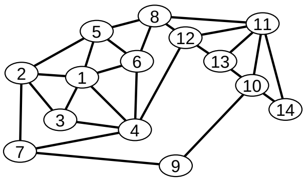
Following the uJaccard similarity and the paradigmatic approach, the results of the graph in figure 3 are shown in table 3.1. we notice that uJaccard similarity provides better information of similarity than Jaccard, this is because uJaccard considers the notion of unilateral similarity. Table 3.1 shows three toy graphs, in which we present a comparison between Jaccard and uJaccard. As show in table 3.1 uJaccard provides a unilateral similarity improving the symmetric similarity Jaccard.

Table 1: uJaccard calculation from figure 3

uJaccard (V1,V7) = 3/5 = 0.600
uJaccard (V7,V1) = 3/7 = 0.428
uJaccard (V7,V12) = 4/7 = 0.571
uJaccard (V12,V7) = 4/4 = 1.000
Jaccard (V1,V7) = 3/9 = 0.333
Jaccard (V7,V1) = 3/9 = 0.333
Jaccard (V7,V12) = 4/7 = 0.571
Jaccard (V12,V7) = 4/7 = 0.571

Table 3.1 shows three toy graphs, in which we present a comparison among Jaccard and uJaccard. As show in table 3.1 uJaccard provide an unilateral similarity improving the symmetric similarity Jaccard.

Table 2: Test uJaccard in toy graphs

	uJaccard sim(2,5)=4/4 sim(5,2)=4/6
	Jaccard sim(2,5)=4/6 sim(5,2)=4/6
	uJaccard sim(7,5)=1/3 sim(5,7)=1/1
	Jaccard sim(7,5)=1/3 sim(5,7)=1/3
	uJaccard sim(2,4)=3/4 sim(4,2)=3/5
	Jaccard sim(2,4)=3/6 sim(4,2)=3/6

3.2 Cut a graph

In graph theory, a cut is a partition of the vertices of a graph into two disjoint subsets. There are many techniques and algorithms to cut a graph, but in some cases there are graphs that are difficult to cut, due to their symmetric distribution of vertices.

It is shown in figure 3.2 that node 1 might belong to cluster {2,3,4} or cluster {5,6}; to resolve this problem we use uJaccard similarity measure to find the similarity of node 1 to other nodes. Table 3 shows that similarities from node 1 to other nodes 1 level deep are the same, so we could not allocate node 1 to a particular cluster. Table 3 also shows that similarities from node 1 to other nodes 2 levels deep, in which uJaccard(1,3) has a strong similarity over the rest. We could conclude that node 1 belong to cluster {2,3,4}.

In figure 3.2 also node 1 might belong to cluster {2,3,4} or cluster {5,6,7} or cluster {8,9,10,11}; this is where uJaccard comes in, being able to solve this problem. Table 4 shows result of similarities from node 1 to all other nodes on the network in different levels deep. cluster {8,9,10,11} presents the highest number of strong similarities, therefor we can conclude that node 1 belongs to cluster {8,9,10,11}.

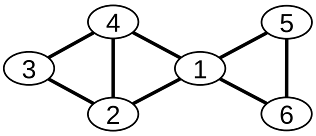


Figure 3: Toy graph.

Table 3: Cut a graph 3.2 using uJaccard

1 level deep		2 levels deep	
uJaccard(1,4)	1/4	uJaccard(1,2)	1/4
uJaccard(1,2)	1/4	uJaccard(1,3)	2/4
uJaccard(1,5)	1/4	uJaccard(1,4)	1/4
uJaccard(1,6)	1/4	uJaccard(1,5)	1/4
-	-	uJaccard(1,6)	1/4

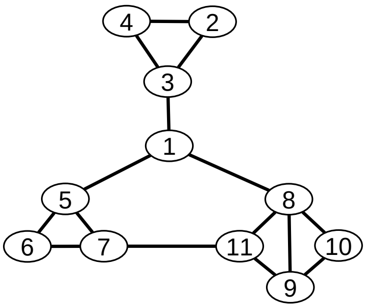


Figure 4: Toy graph.

3.3 Social Network

We tested uJaccard against two social network graphs; the first is the coauthorship network of scientists [9] the second is the network of Hollywood's actors[1].

The first network is the coauthorship network of scientists working on network theory and experiments, compiled by M. Newman [9]. We want to find the top scientists that Newman is similar to or that have paradigmatic similarity. As shown in table 5 the 3 most of Newman's paradigmatic similar scientists are Callaway, Strogatz and Holme. On the

[1] The Internet Movie Database: ftp://ftp.fu-berlin.de/pub/misc/movies/database/

Table 4: Cut a graph 3.2 using uJaccard

2 levels deep		3 levels deep	
uJaccard(1,4)	1/3	uJaccard(1,4)	1/3
uJaccard(1,2)	1/3	uJaccard(1,2)	1/3
uJaccard(1,6)	1/3	uJaccard(1,6)	1/3
uJaccard(1,7)	1/3	uJaccard(1,7)	2/3
uJaccard(1,9)	1/3	uJaccard(1,9)	2/3
uJaccard(1,11)	1/3	uJaccard(1,11)	2/3
uJaccard(1,10)	1/3	uJaccard(1,10)	1/3

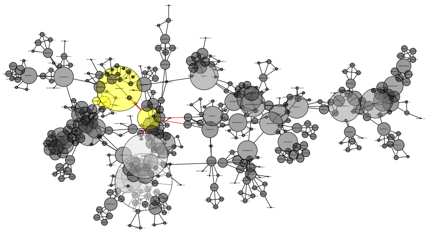


Figure 5: Coauthorship network of scientists, selected nodes belong to scientists Newman, Callaway, Strogatz, Holme.

other hand the top 3 scientists that are similar to Newman are Adler, Aberg and Aharony. uJaccard has been calculated in 2,3 and 4 levels deep away from Newman. Newman is more similar to Strogatz but the most similar scientist to new Newman is Adler and not Strogatz, even that Strogatz most similarity is toward Newman.

For the second network, we created the second social network of Hollywood's actors, we based on The Internet Movie Database (note). We download actors and actresses data, which includes title of movies in which they worked, we also download a list of top 1000 (nota) and top 250 (nota) actors and actresses. The network is composed of nodes representing actors and actresses, and vertices are the movies in which those actors worked together. A node is created for every person, with their names as the key, when two people are in the same movie; a vertex is created between their nodes. The first network presents 1000 top actors and actresses who also work in 41,719 movies with a total 113,478 edges. The second network presents 250 actors and actresses who work in 15,831 movies with a total of 14,096 edges. For this test we remove duplicated edges.

- From a given *actor A*
- We search for actors that *actor A* is similar to
- From the *actor A*'s similar actor list we get the most similar *actor B*
- We search for actors that *actor B* is similar to
- This is done to analyse if *actor A* and *actor B* are reciprocally similar
- Then we look for actors that are most similar to *actor A*
- We do this on the network top 250 actors and top 1000 actors.

Table 5: uJaccard calculation from 3.3, in search paradigmatic scientists to Newman

2 levels deep		3 levels deep		4 levels deep	
Scientist Newman is similar to:					
Callaway	0.15	Strogatz	1.63	Strogatz	8.25
Strogatz	0.15	Holme	1.59	Callaway	7.85
Watts	0.15	Kleinberg	1.59	Watts	7.81
Hopcroft	0.11	Sole	1.59	Kleinberg	7.18
Scientists that are similar to Newman:					
Adler	0.33	Aberg	0.50	Aberg	2.50
Aharony	0.33	Adler	0.66	Adler	14.0
Aleksiejuk	0.50	Aharony	0.66	Aharony	14.0
Ancelmeyers	0.66	Alava	0.50	Alava	1.00
Araujo	0.33	Albert	0.10	Albert	0.50

The results of the search on the network of top 250 actors and top 1000 actors, using uJaccard and the paradigmatic approach are presented in tables 6 and 7. In table 6 we focus in *Tom Cruise*, we found that *Tom Cruise* is most similar to *Julia Roberts* but *Julia Roberts* is most similar to *John Travolta*, *Tom Cruise* is third in *Julia Roberts'* similarity list. clearly there is not a symmetric similarity among *Julia Roberts* and *Tom Cruise*. Moreover *Julia Roberts* is not the most similar toward *Tom Cruise*, the most similar towards *Tom Cruise* is *Heath Ledger*. Hence this confirm that uJaccard helps to identify similarities, particularly asymmetric similarities. Table 6 also shows similar scenario among *Tom Cruise*, *Tom Hanks* and *Joan Allen* in the network of top 250 actors and actresses, this confirm the usability of uJaccard.

In Table 7 we use the network of top 250 actors and actresses and we focus on *Anthony Quinn* and *Jack Nicholson*. We start by searching for actors that *Anthony Quinn* is sim-

Table 6: uJaccard similarity among top 1000 and top 250 actor and actresses, searching paradigmatic similar actor

Top 1000 actors, cruise tom is similar to:			
		roberts julia	
roberts julia	0.405	travolta john	0.418
hanks tom	0.401	hanks tom	0.412
jackson samuel	0.399	jackson samuel	0.407
douglas michael	0.397	cruise tom	0.399
eastwood clint	0.393	spacey kevin	0.399
Top 250 actors, cruise tom is similar to:			
		hanks tom	
hanks tom	0.430	douglas michael	0.425
douglas michael	0.420	cruise tom	0.421
eastwood clint	0.420	jackson samuel	0.418
spacey kevin	0.413	travolta john	0.414
jackson samuel	0.410	spacey kevin	0.411
Who is similar to cruise tom:			
top 1000 actors		top 250 actors	
ledger heath	0.490	allen joan	0.496
bacon kevin	0.488	balk fairuza	0.495
crowe russell	0.482	bello maria	0.488
gibson mel	0.482	collins pauline	0.487
benigni roberto	0.482	aiello danny	0.487

ilar to, then we search for actors that are most similar to *Anthony Quinn* in 1 and 2 levels deep. We notice that *Anthony Quinn* is most similar to *Tom Hanks* but most similar actor to *Anthony Quinn* is *Antonio Banderas*, while *Anthony Quinn* is the 153th most similar for *Tom Hanks. Antonio Banderas* is most similar to *Samuel Jackson* and not to *Anthony Quinn*, while *Anthony Quinn* is the 53th most similar for *Antonio Banderas.* Therefore we could conclude that *Anthony Quinn* and *Tom Hanks* are not symmetric similar rather they are asymmetric similar. Table 7 also shows similar scenario for *Jack Nicholson.*

Table 7: uJaccard similarity among top 250 actor and actresses, searching paradigmatic actor, in 2 and 3 levels deep

Top 250 actors, are similar to:			
quinn anthony		nicholson jack	
hanks tom	0.451	hanks tom	0.436
jackson samuel	0.443	eastwood clint	0.429
lemmon jack	0.443	travolta john	0.417
cruise tom	0.435	williams robin	0.417
de niro robert	0.435	douglas michael	0.414
Who are similar to quinn anthony:			
1 level deep		2 levels deep	
banderas antonio	0.366	banderas antonio	21.866
bardem javier	0.350	benigni roberto	20.758
martin steve	0.333	burns george	20.565
goodman john	0.320	baldwin alec	20.413
allen woody	0.285	mcqueen steve	20.244
Who are similar to nicholson jack:			
1 level deep		2 levels deep	
greene graham	0.484	banderas antonio	41.477
bronson charles	0.482	bacon kevin	41.133
brody adrien	0.480	baldwin alec	41.0862
bale christian	0.476	bronson charles	40.982
baldwin alec	0.474	benigni roberto	40.948

4. CONCLUSION

A key assumption of most models of similarity is that a similarity relation is symmetric. The symmetry assumption is not universal, and it is not essential to all applications of similarity. The need for asymmetric similarity is important and central in Information Retrieval and Graph Data Networks. It can improve current methods and provide an alternative point of view.

We present a novel asymmetric similarity, Unilateral Jaccard Similarity (uJaccard), where the similarity among A and B is not same to the similarity among B and C, *uJaccard(A,B) != uJaccard(B,A)*; this is based on the idea of paradigmatic association. In comparison to Tversky [13] our approach uJaccard does not need a stimulus bias, whereas in the case of Tversky human judgement is needed.

We present a series of cases in which we confirmed its usefulness and we validated uJaccard. We could extend uJaccard to include weights to improve the asymmetry, we could also use uJaccard and the paradigmatic approach to cluster Graph data Networks. These are tasks in which we are working on.

In conclusion, the proposed uJaccard similarity proved to be useful despite its simplicity and the few resources used.

5. ACKNOWLEDGMENTS

This research was funded in part by ”Fondo para la Innovación, Ciencia y Tecnología - Peru” (FINCyT). We thank the Universidad Católica San Pablo for supporting this research.

6. REFERENCES

[1] D. Bridge. Defining and combining symmetric and asymmetric similarity measures. In B. Smyth and P. Cunningham, editors, *Advances in Case-Based Reasoning (Procs. of the 4th European Workshop on Case-Based Reasoning)*, LNAI 1488, pages 52–63. Springer, 1998.

[2] F. De Saussure and W. Baskin. *Course in general linguistics.* Columbia University Press, 2011.

[3] R. Duda, P. Hart, and D. Stork. Pattern classification 2nd ed., 2001.

[4] C. Fellbaum. Wordnet and wordnets. In A. Barber, editor, *Encyclopedia of Language and Linguistics*, pages 2–665. Elsevier, 2005.

[5] E. W. Holman. Monotonic models for asymmetric proximities. *Journal of Mathematical Psychology*, 20(1):1–15, 1979.

[6] S. Jimenez, C. Becerra, and A. Gelbukh. Soft cardinality: A parameterized similarity function for text comparison. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 449–453. Association for Computational Linguistics, 2012.

[7] L. Lee, F. C. Pereira, C. Cardie, and R. Mooney. Similarity-based models of word cooccurrence probabilities. In *Machine Learning.* Citeseer, 1999.

[8] F. Lorrain and H. C. White. Structural equivalence of individuals in social networks. *The Journal of mathematical sociology*, 1(1):49–80, 1971.

[9] M. E. Newman. Finding community structure in networks using the eigenvectors of matrices. *Physical review E*, 74(3):036104, 2006.

[10] R. M. Nosofsky. Stimulus bias, asymmetric similarity, and classification. *Cognitive Psychology*, 23(1):94–140, 1991.

[11] M. Sahlgren. The word-space model: Using distributional analysis to represent syntagmatic and paradigmatic relations between words in high-dimensional vector spaces. 2006.

[12] R. N. Shepard. Representation of structure in similarity data: Problems and prospects. *Psychometrika*, 39(4):373–421, 1974.

[13] A. Tversky. Features of Similarity. In *Psychological Review*, volume 84, pages 327–352, 1977.

[14] J. Weeds and D. Weir. Co-occurrence retrieval: A flexible framework for lexical distributional similarity. *Computational Linguistics*, 31(4):439–475, 2005.

[15] D. R. White and K. P. Reitz. Graph and semigroup homomorphisms on networks of relations. *Social Networks*, 5(2):193–234, 1983.