Métricas de similitud para la identificación de especies de plantas a través de vectores de características

Ruiz Castilla José Sergio
Universidad Autónoma del Estado de México
Centro Universitario UAEM Texcoco, Texcoco, Estado de México, México
jsergioruizc@gmail.com

Juárez Rodríguez Ricardo Rodrigo
Universidad Autónoma del Estado de México
Centro Universitario UAEM Texcoco, Texcoco, Estado de México, México
guillermus@prodigy.net.mx

Cervantes Canales Jair
Universidad Autónoma del Estado de México
Centro Universitario UAEM Texcoco, Texcoco, Estado de México, México
chazarra17@gmail.com

Trueba Espinosa Adrián
Universidad Autónoma del Estado de México
Centro Universitario UAEM Texcoco, Texcoco, Estado de México, México
atruebae@gmail.com

***Resumen*— En el Instituto Botánico de la UNAM existe un repositorio de 21767 especies de plantas, las cuales han sido recolectadas y clasificadas en familias durante años. El proceso de identificación y clasificación está a cargo del Dr. José Luis Villaseñor Ríos en calidad de "curador". Se han acumulado miles de ejemplares en espera de ser identificados y clasificados, para lo anterior el Dr. Villaseñor ha creado una lista de 150 características que pueden tener o no las familias de cada planta, asignando un "0" cuando no posee la característica y un "1" cuando si la posee. Las 150 características forman un vector de 0´s y 1´s que se usaron para estudiar su similitud con otras familias mediante las métricas de *Simple Matching Coefficient* (SMC), *Jaccard* y *Rogger & Tanimoto*. Se han obtenido resultados donde el método SMC arroja mejores resultados permitiendo identificar una planta y su similitud con otras especies.**

***Palabras clave*—identificación, métricas de similitud, plantas, vector de caracteres.**

I. INTRODUCCIÓN

En México, en el Instituto Botánico de la UNAM se coleccionan, identifican, clasifican y conservan 21767 especies de plantas. La clasificación la coordina y valida el Dr. Villaseñor Ríos José Luis en calidad de "curador", el cual supervisa el proceso de identificación, y clasificación de las especies, y por otra parte, se tienen almacenadas miles de muestras de plantas recolectadas en espera de ser identificadas y clasificadas. Se debe considerar que, la identificación taxonómica se define como el proceso de nombrar o catalogar un espécimen dentro de un sistema de clasificación previa. Comprende tres actividades principales que son: clasificación, nomenclatura e identificación [1]. La clasificación es el proceso de asignar un espécimen o grupo de especímenes dentro de una categoría taxonómica, o sea un taxón. En la nomenclatura se le asigna un nombre a las diferentes categorías acorde al Código Internacional de Nomenclatura Botánica [2]. Por último, en la identificación se determina dónde debe estar ubicado un espécimen dentro de este sistema de clasificación internacional botánico.

Anteriormente el mencionado Instituto desarrolló una herramienta para ayudar en la tarea de clasificación llamada GENCOMEX: a Computerized Key to Identify the Genera of Asteraceae Of México [2], desarrollada en 1998 en el lenguaje de programación denominado "Pascal" por lo que no es compatible con los Sistemas Operativos actuales, por lo que lleva un rezago de la herramienta de decenas de años. Por lo anterior, se propuso una herramienta de software para apoyar a los expertos en botánica en la actividad de identificación de las familias de plantas en cuestión.

Se han definido 150 características para cada especie para identificar y clasificar por clase y familia, con las cuales se ha construido un vector de caracteres donde un cero significa ausencia de la característica y un 1 cuando existe la característica. Cada vector de características representa la cadena de identificación de cada especie. Cuando se obtienen una nueva muestra es necesario conocer y registrar sus características para identificar a que especie corresponde. Cuando una muestra de planta es recolectada más de una vez, se registra añadiendo la fecha y lugar de recolección para un posterior mapeo y estudio del comportamiento de la especie en cuestión.

Se han desarrollado algoritmos usando las técnicas de distancia *Simple Matching Coefficient* (SMC), Jaccard, y la de Rogger & Tanimoto, para conocer el grado de similitud que guarda una cadena de caracteres con otra, por lo cual, en este caso se usaron para identificar una especie, dentro de una colección de especies ya existente. Se han implementado tres algoritmos y generaron resultados que al compararse ha sido posible determinar la mejor técnica. Primero, fue necesario conocer la distancia que indica la cantidad de caracteres que son diferentes en las cadenas comparadas. Cuando no existen caracteres diferentes las cadenas son iguales y la distancia es 0, cuando un carácter es diferentes la distancia es 1 y así

sucesivamente. Una vez conocida la distancia, así como las similitudes y disimilitudes se busca el grado de similitud.

El proyecto tiene como propósito desarrollar una herramienta que permita la identificación de especies de plantas para apoyar a los expertos en botánica en la tarea de identificar especies de plantas a partir de las características principalmente de las hojas.

Este trabajo de investigación tiene como objetivo identificar la mejor técnica de similitud de entre SMC, Jaccard y Rogger & Tanimoto para la identificación de plantas con sus vectores de características.

Se revisó el trabajo de investigación de García que presenta métricas de similitud para búsqueda aproximada con técnicas de *Aproximate String Matching* (ASM) para detectar patrones de cadenas que permiten errores. ASM mide las distancias con: Levenshtein, Haming, Jaro, Markov entre otros métodos, obteniendo resultados de similitud de dos algoritmos de programación dinámica: con el cálculo de distancia de Levenshtein y paralelización por bits. Por lo anterior los mencionados métodos permiten la comparación de dos cadenas y saber el grado de similitud. Destaca además que el grupo de cadenas existente es el alfabeto contra el cuál se compara un patrón, que la cadena de caracteres está formada por caracteres o por 0's y 1's, que cuando las cadenas son iguales se logra una búsqueda exacta, mientras que al existir diferencias es necesario buscar la distancia y la métrica de similitud. El objetivo que se buscó consiste en identificar errores y proceder a una adición, eliminación o sustitución de caracteres para hacer una corrección de textos. Argumenta además, que las métricas de distancia de similitud permiten aplicaciones en: bio-informática, reconocimiento de patrones, procesamiento de señales, comparación de archivos y corrección de texto, minería de datos y en bases de datos. Finalmente cabe señalar que ASM procesa dos cadenas de diferente tamaño mientras que esta investigación se trabaja con cadenas de igual tamaño [3]. Es posible observar un ejemplo de la distancia de Levenshtein en la Tabla 1.

Tabla 1. Ejemplo de la Distacia de Levenshtein[3].

Universalidad vs aniversario												
U	N	I	V	E	R	S	A	L	I	D	A	D
x	|	|	|	|	|	|	|	x	|	x	x	x
A	N	I	V	E	R	S	A	R	I	O	-	-

En el ejemplo de Levenshtein la distancia es 5, porque es la cantidad de caracteres diferentes y sería necesaria la sustitución de cinco caracteres para que las cadenas sean iguales. Lo anterior cuando el objetivo es una corrección automática, de lo contrario solo podría buscarse la métrica de similitud.

En otro trabajo de investigación de Cohen y otros hacen una comparación de nombre y registros en una base de datos de un censo a través de algoritmos de *String Matching*. En dicha investigación se aplicaron diversas técnicas obteniendo los siguientes resultados de la Tabla 2.

Tabla 2. Resultados de métricas de cadenas [4].

	MaxF1	AvgPrec	
SFS	0.528	0.0357	
TFIDF	0.518	0.369	
Jaccard	0.567	0.402	
L2 JaroWinkler	0.746	0.770	
SoftTFIDF	0.685	0.782	
Jaro-Winkler	0.648	0.703	
Jaro	0.687	0.731	
NaiveAvgOverlap	0.697	0.731	
AvgOverlap	0.701	0.736	
Levenshtein	0.832	0.901	
Jaro	0.728	0.789	Recortado
Scaled Levenstein	0.851	0.0930	Recortado
Levenshtein	0.865	0.925	Recortado

De acuerdo a los resultados el método con menor precisión ha sido SFS, mientras que el mejor es el de Levenshtein concluyendo que las técnicas de Jaro y Levenshtein son eficaces en el cálculo de la métrica de similitud de dos cadenas de diferente tamaño [4].

II. Método utilizado

Se tomaron en cuenta los datos proporcionado por el Dr. Villaseñor Ríos de los vectores de 0's y 1's de los cuales, se tomaron 10 vectores (del conjunto total) con 150 características correspondientes a 10 especies, precisando que son cadenas de bits del mismo tamaño. Luego, se usó la especie *Orchidaceae* como la especie a identificar y se comparó consigo misma y con las otras nueve especies para procesar la identificación. Las técnicas algorítmicas de métricas de similitud que se usaron, son: de SMC, Jaccard, y de Rogger & Tanomoto, considerando que las cadenas a comparar son del mismo tamaño.

En base a los procedimientos establecidos por el Dr. José Luis Villaseñor Ríos se determina como se identifican los grupos de Clases y Familias en México. La primera gran división es entre 2 clases que son *Liliopsida* (4500 especies) y *Magnoliopsida* (17500 especies). Dentro de la clase *Liliopsida* se encuentran 2 grandes familias que son la Orchidaceae que posee cerca de 1500 especies y la *Poaceae* con 1000 especies aproximadamente. Por la parte de la clase de *Magnaliopsida* que es el conjunto mayor donde se encuentran 2 grandes familias que son la *Asteraceae* con cerca de 9000 especies y la *Fabaceae* con 2000 especies aproximadamente [6].

A. Disposición de la información fuente

La información proporcionada por el Instituto de Botánica de la UNAM está dispuesta primero por una lista de 21767 especímenes de plantas. Una planta se clasifica en clase, familia, género y especie [1], de las cuales se muestra la estructura de la lista para identificar la familia a la cual posee una planta. Ver Tabla 3.

Tabla 3. Lista de plantas clasificada hasta especie [2].

	Clase	Familia	Género	Especie
1	Liliopsida	Alismataceae	Echinodorus	andrieuxii
2	Liliopsida	Alismataceae	Echinodorus	berteroi
3	Liliopsida	Alismataceae	Echinodorus	cordifolius
4	Liliopsida	Alismataceae	Echinodorus	grandiflorus
5	Liliopsida	Alismataceae	Echinodorus	nymphaeifolius
6	Liliopsida	Alismataceae	Echinodorus	paniculatus
7	Liliopsida	Alismataceae	Echinodorus	tenellus
8	Liliopsida	Alismataceae	Echinodorus	virgatus
9	Liliopsida	Alismataceae	Helanthium	Bolivianum
...	...	...	...	...
21776	Magnoliopsida	Zygophyllaceae	Viscainoa	Geniculata

Se ha creado una lista de 150 características que pueden o no poseer cada especie que se muestra en la Tabla 4.

Tabla 4. Características de las familias [2].

	Características
1	Plantas leñosas (árboles o arbustos)
2	Plantas herbáceas (anuales o perennes, incluyendo sufrútices)
3	Bejucos o plantas escandentes
4	Plantas acuáticas o subacuáticas
5	Plantas epífitas
6	Plantas parásitas o saprófitas
7	Plantas con jugo lechoso (látex)
8	Plantas con jugo acuoso (no lechoso)
9	Plantas aromáticas o resinosas (en corteza, ramas u hojas)
10	Plantas con zarcillos
...	...
150	Vegetación acuática

B. Caracterización de las hojas

De igual forma y siguiendo con el procedimiento establecido por el Dr. Villaseñor, se usó la convención de 1's cuando posee la característica y 0's si hay ausencia de dicha característica, dentro de las 150 posibles características, resultando un vector de 150 valores con 1's y 0's. La Tabla 5 muestra las características de cada familia.

Tabla 5. Familias con sus características [6].

Clase	Familia	2	3	4	5	...	150
Liliopsida	Agavaceae	1	0	0	0	...	0
Liliopsida	Alismataceae	1	0	1	0	...	1
Liliopsida	Alliaceae	1	0	0	0	...	1
Liliopsida	Aloeaceae (I)	1	0	0	0	...	0
Liliopsida	Alstroemeriaceae	1	1	0	0	...	0
Liliopsida	Amaryllidaceae	1	0	0	0	...	1

C. Definición de patrones

Los patrones se establecieron acorde a la Tabla 6, la cual se conforma en una matriz con dos dimensiones. La primera dimensión correspondiente a la fila para los valores de la cadena 1, mientras que la primera columna corresponde a los valores de la cadena 2. La combinación de valores permite obtener cuatro variables en este caso, a, b, c, y d, con los valores 11, 10, 01 y 00. Lo anterior se usó en las fórmulas 1, 2, y 3 para obtener los resultados buscados.

Tabla 6. Tabla de patrones [5].

		Cadena 1	
		1	0
Cadena 2	1	**11** **a**	**10** **b**
	0	**01** **c**	**00** **d**

La fórmula de SMC considera las similitudes 11 y 00 dividida entre las similitudes 11, 00 y las disimilitudes 10 y 01. Fórmula 1.

$$SMCij = \frac{a+d}{a+b+c+d} \qquad (1)$$

Dónde i es la cadena 1 y j es la cadena 2.

En el caso de *Jaccard* considera solo las similitudes existentes, es decir 11, cuando las dos cadenas contienen la características, dividido entre similitudes 11 y las disimilitudes, pero no incluye similitudes 00. Fórmula 2.

$$Jacard\,(X1, X2) = \frac{a}{a+b+c} \qquad (2)$$

Dónde X1 es la cadena 1 y X2 es la cadena 2.

En el caso de la Fórmula de Rogger & Tanimoto divide las similitudes 11 y 00 entre las similtudes 11, 00 y el doble de las disimiltudes 10 y 01. Fórmula 3.

$$Rogger\&Tanimoto = \frac{a+d}{a+d+2(b+c)} \qquad (3)$$

Se desarrolló un programa para obtener las métricas de similitud de SMC, Jaccard, y Rogger & Tanimoto y se muestra en el código 1.

Se usó un vector de 0's y 1's de características de una planta para compararse con otros 9 vectores de 0's y 1's elegidos para la experimentación, se hicieron las corridas necesarias para obtener resultados con diferentes distancias.

D. Diseño del algoritmo

```
main program

   Family patron("Orchidaceae");
   DBFamilies dbFamilies("Orchidaceae", "Poaceae",
      "Liliaceae", "Lacandoniaceae", "Agavaceae",
      "Asteraceae", "Fabaceae","Magnoliaceae",
      "Santalaceae", "Tiliaceae");

   for_each familiy : dbFamilies

      write patron name;
      write family name;
      write distance (patron, familie);
      var a get Number Of A's (patron, family);
      var b get Number Of A's (patron, family);
      var c get Number Of A's (patron, family);
      var d get Number Of A's (patron, family);
      write SimSMC(a, b ,c, d);
      write SimJaccard(a, b ,c);
      write SimRoggerTanimoto(a, b, c, d);

   end for_each

   SimSMC(a, b, c, d) {
      (a + b) / (a + b + c + d);
   }

   SimJaccard(a, b, c) {
      a  / (a + b + c );
   }

   SimRoggerTanimoto(a, b, c, d) {
      (a + b) / (a + b + 2*(c + d));
   }

end_main program
```

III. Resultados

Se probaron los tres algoritmos y se obtuvieron resultados que se muestran en la Tabla 5 donde se puede observar que a medida que aumenta la distancia, disminuye el grado de similitud. Por otro lado se puede observar que las tres técnicas arrojan valores diferentes, pero consistentes. La técnica con valores menos preciso es la de Rogger & Tanimoto, mientras que la que arroja los mejores resultados es SMC, ver Tabla 7.

En este caso si la distancia es 0 significa que las cadenas son iguales y la similitud es 100%, por lo tanto si el experto en botánica introduce las características completas y correctas, logrará una "búsqueda exacta" [3], en caso contrario obtendrá un porcentaje que tiende a 100% a medida que exista similitud con una especie almacenada en la aplicación.

Table 7. Resultados de grados de similitud.

Patrón	Familia	Distancia	A 11	B 10	C 01	D 00	SMC	Jaccard	Rogger & Tanimoto
Orchidaceae	Orchidaceae	0	99	0	0	51	1.0	1.0	1.0
Orchidaceae	Agavaceae	33	76	23	10	41	0.78	0.697	0.639
Orchidaceae	Poaceae	43	73	26	17	34	0.713	0.629	0.554
Orchidaceae	Tiliaceae	50	71	28	22	29	0.666	0.586	0.5
Orchidaceae	Asteraceae	51	73	26	25	26	0.66	0.588	0.492
Orchidaceae	Fabaceae	51	76	23	28	23	0.66	0.598	0.492
Orchidaceae	Liliaceae	68	53	46	22	29	0.546	0.438	0.376
Orchidaceae	Magnoliaceae	73	48	51	22	29	0.513	0.396	0.345
Orchidaceae	Lacandoniaceae	87	28	71	16	35	0.42	0.243	0.265
Orchidaceae	Santalaceae	93	30	69	24	27	0.38	0.243	0.234

Los resultados nos indican el grado de similitud considerando la distancia y los resultados tienen el mismo comportamiento en cada técnica, solo que cada una con diferente valor respecto al 100%. En la Fig.1. Se puede observar que SMC se mantiene por encima de Jaccard y Rogger & Tanimoto. En la práctica cuando el botánico no está seguro de la identificación de la especie es cuando acude al experto "curador" para que emita su dictamen y determinar de qué especie se trata.

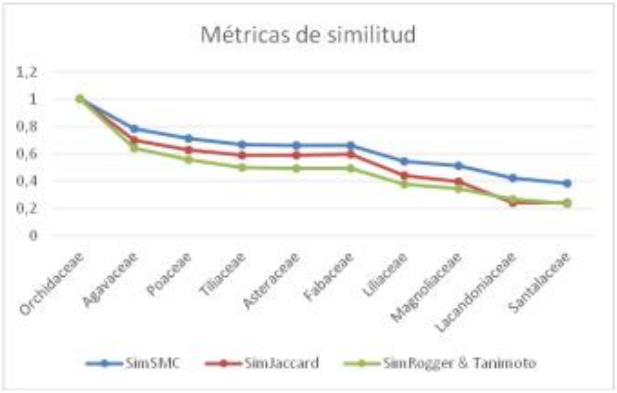

Fig. 1. Comportamiento de las técnicas utilizadas.

IV. Conclusiones

Se concluye que con los métodos de SMC, Jaccard y Rogger & Tanimoto permiten identificar especies de plantas usando el vector de características con 0's y 1's y que el método más eficiente ha sido el de SMC.

Que es posible apoyar a los alumnos e investigadores de botánica en la identificación de especies al marcar una a una de las características de una planta a identificar con mayor eficacia. Con lo anterior se crea un vector de 0's y 1's y con la aplicación propuesta es posible un resultado exacto, o bien, en caso de que no se incorporen algunas características, la distancia determinará a cuál especie se parece.

Una ventaja importante de la herramienta propuesta es que el usuario puede portar la aplicación en un dispositivo móvil debido a la eficiencia de representación, manejo y

}

procesamiento de información, y en trabajo de campo se puede hacer la caracterización y obtener resultados de inmediato.

A. Trabajos futuros

Como parte del mismo proyecto, se está trabajando otro proyecto que consiste en la posibilidad de tomar una fotografia, caracterizar la hoja de la planta y hacer una identificación a través de procesamiento de imágenes de manera automática.

REFERENCIAS

[1] J. L. Villaseñor, "La familia Asteraceae en México", in Revista de la Sociedad Mexicana de Historia Natural, 1993, pp. 117-124.

[2] J. L. Villaseñor R, "GENCOMEX a computarizad key to identify the genera of asteraceae of México", in Asociación de Biólogos de la Computación A. C, 1998.

[3] J. F. García, "Métricas de Similitud para Búsqueda Aproximada", in Revista Tecnológica de Ingeniería en Sistemas, Vol. 6, no. 2, Ed. Políticas Editoriales, 2007, pp. 15-27.

[4] W. W. Cohen, P. Ravikunar, P. Fienberg S.," A comparison of String Metrics for Matching Names and Records", in Kdd workshop on data cleaning and object consolidation, Vol. 3, 2003, pp 73-78.

[5] R. Valenzuela, D. Blanca, "Marco orientado a objetos para cálculos de similitud", in Centro Nacional de Investigación y Desarrollo Tecnológico, Departamento de Ciencias Computacionales, no, 28, 2012.

[6] R. R. Juárez H, J. S. Ruiz C, J. Cervantes C., F. García L., "Reconocimiento de patrones para la identificación de clase y familia de plantas a partir de sus caracteres", in Cuerpos Académicos de la DES Oriente en búsqueda de la implementación de la ciencia y la tecnología 2014, 2015, pp. 402-417.