Clasificación de un conjunto de diagramas de clases mediante QSI (QSI)

F. J. Cuberos[1], J. A. Ortega[2], M. Genero[3], M. Piattini[3]

[1]Dpto.de Planificación. Radio Televisión de Andalucía.
Ctra.San Juan-Tomares km.1,3.S.J.Aznalfarache –Sevilla
fjcuberos@rtva.es
[2]Dpto.de Lenguajes y Sistemas Informáticos.Universidad de Sevilla
Avda.Reina Mercedes s/n - Sevilla
ortega@lsi.us.es
[3]Departmento De Informática, Universidad de Castilla-La Mancha
Paseo de la Universidad, 4 - Ciudad Real
{Marcela.Genero, Mario.Piattini}@uclm.es

Resumen. El objetivo de este trabajo es aplicar el índice cualitativo de similitud (QSI) para agrupar un conjunto de diagramas de clases según su complejidad estructural, medida a través de un conjunto de métricas cuantitativas. Además se comparan los resultados obtenidos en un trabajo previo, en el que se utilizó otra técnica de clustering jerárquico similar al que se usa en los "Repertory Grids".

1. Introducción

En primer lugar en esta introducción explicaremos cual fue la motivación que llevó a la definición del QSI y a continuación describiremos el problema sobre el cual lo aplicaremos posteriormente.

1. 2 Introducción al QSI

En la bibliografía se han presentado diferentes aproximaciones para comparar series temporales. La mayoría proponen la creación de un índice con un pequeño conjunto de valores extraídos de los datos originales.

Para la generación del índice se ha optado mayoritariamente por dos enfoques diferentes: el realizar una transformación de los valores de la serie temporal a un espacio de menor dimensión y la de reducir directamente los datos originales de la serie temporal seleccionando un subconjunto de ellos.

La transformación a espacios de menor dimensión se realiza por medio de utilizar la Transformada Discreta de Fourier, como en [1], [24], [10], [23], [8], [12] y [13], o Discrete Wavelet Transform, como la Transformada de Haar, que es usada en [4] y [11].

La aproximación de reducir directamente los datos de las series, seleccionando un subconjunto de ellos, se presenta en [14],[15],[16], [27], [
18] y [19].

Un tercer grupo de trabajos se basan en la aplicación del Dynamic Time Warping, o *DTW*, algoritmo muy conocido por su propiedad de encontrar coincidencias aunque no estén alineadas en el tiempo. Algunos de estos trabajos aplican previamente técnicas de reducción de los datos originales y entre ellos encontramos [15],[17],[26],[5],[20],[21],[25] y [22].

El algoritmo de la subsecuencia común máxima, o *LCS*, puede considerarse un caso particular del *DTW*, manteniendo todas sus características. Se han diseñado varias soluciones para la aplicación eficiente de este algoritmo que son revisadas en [7].

El índice QSI es novedoso al considerar los valores de los atributos desde una perspectiva cualitativa, por medio de su conversión a etiquetas cualitativas, y a su posterior comparación con un algoritmo de similitud de cadenas de caracteres.

1.2 Descripción del problema

Con el objetivo de averiguar si existe correlación entre un conjunto de métricas definidas para medir la complejidad estructural de los diagramas de clases UML, y la mantenibilidad de dichos diagramas se realizó un experimento controlado del que se obtuvieron los datos mostrados en la tabla 1.

Tabla 1. Datos obtenidos en un experimento controlado [9].

	NC	NA	NM	NAssoc	NAgg	NDep	NGen	NaggH	NGenH	MaxDIT	MaxHAgg	Tiempo
D1	7	11	22	1	0	0	5	0	1	2	0	2,696
D2	8	12	31	1	6	0	1	1	1	1	2	3,043
D3	3	17	24	2	0	0	0	0	0	0	0	2,913
D4	10	12	21	15	3	0	0	2	0	0	1	4,087
D5	9	19	29	3	3	0	3	3	1	2	1	3,345
D6	7	16	7	6	0	0	0	0	0	0	0	2,739
D7	23	33	66	4	5	2	16	2	3	3	3	4,87
D8	20	30	65	6	5	0	14	4	3	3	2	4,826
D9	23	65	80	20	3	2	3	3	1	2	3	7,087

Las 9 filas representan a cada uno de los 9 diagramas de clases que se le entregaron a los sujetos. Las 11 primeras columnas representan los valores de las variables dependientes (métricas) y la última representa la media de los valores de la variable dependiente, en este caso el tiempo promedio de mantenimiento. Dicho tiempo se obtuvo como resultado del las tareas llevadas a cabo en el experimento, en el que los sujetos debían modificar los diagramas de clases según nuevos requerimientos y anotar el tiempo empleado en tales modificaciones. Este experimento se realizó con profesores del área de ingeniería de software y alumnos de quinto año.

Para encontrar diferentes prototipos que agrupen los diagramas de clases según su complejidad se utilizó en [9] un clustering jerárquico similar al que se usa en los "Repertory Grids" [3]. En dicho clustering se encontraron tres prototipos que agrupan los diagramas en Complejidad baja, media y alta (ver figura 1).

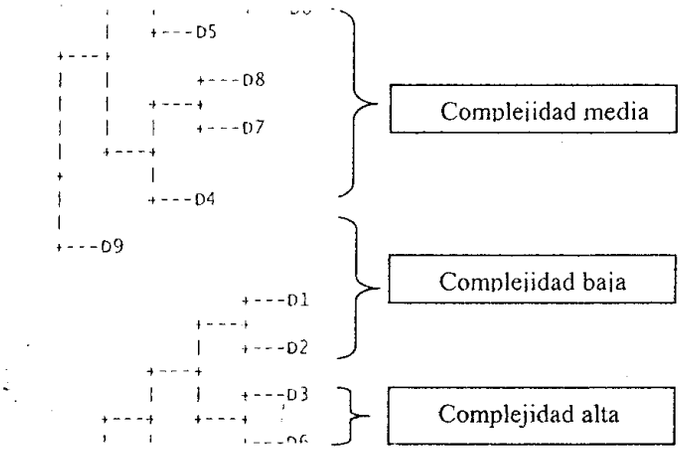

Fig. 1. Dendograma obtenido con los "Repertory Grids" [9]

Lo que pretendemos en este trabajo es comparar el clustering obtenido previamente (ver figura 1) con el que podemos obtener aplicando el QSI.

Este trabajo se organiza de la siguiente manera: La descripción del QSI se presenta en la Sección 2. Los resultados obtenidos al aplicar el QSI a los datos mostrados en la tabla 1 se detallan en la Sección 3, y la comparación con los resultados obtenidos previamente se realiza en la Sección 4. Y por último en la Sección 5 se presentan las líneas de trabajo futuro.

2. Índice Cualitativo de Similitud (QSI)

La idea de este índice es la inclusión de conocimiento cualitativo en la comparación de series temporales. Se propone una medida basada en la coincidencia de etiquetas cualitativas que representan la evolución de los valores de las series. Cada etiqueta representa un rango de valores que pueden asumirse como similares desde una perspectiva cualitativa. Diferentes series con una evolución cualitativamente similar producen la misma secuencia de etiquetas.

Es interesante resaltar que las series temporales se suponen libres de ruido entre dos muestras y con una evolución lineal y monotónica.

Sea $X=\langle x_0,...,x_f\rangle$ una serie temporal. Nuestro acercamiento se aplica en tres pasos.Primero, se realiza una normalización de los valores de X, obteniéndose $\%=\langle \%_0,\mathrm{K},\%_f\rangle$. Utilizando esta serie, se calcula la serie de diferencias

$X_D = \langle d_0, \mathrm{K}, d_{f-1} \rangle$, que se traduce en una cadena $S_X = \langle c_1, \mathrm{K}, c_{f-1} \rangle$. La similitud entre dos series se calcula mediante la comparación de las dos cadenas obtenidas por medio del algoritmo LCS. El resultado se usa como medida de similitud de las series originales.

Aunque el índice QSI está definido para series temporales, en el presente documento vamos a realizar una aplicación del mismo sobre objetos de múltiples atributos. Para ello basta con considerar la lista de los valores de los atributos como una serie temporal.

2.1 Normalización

Se realiza una normalización de los valores originales en el intervalo [0,1]. Esta normalización permite la comparación de series temporales con diferentes escalas cuantitativas.

Sea $\mathcal{X}' = \langle \mathcal{X}'_0, \mathrm{K}, \mathcal{X}'_p \rangle$ la serie normalizada obtenida desde X.

Sea $X_D = \langle d_0, \mathrm{K}, d_{f-1} \rangle$ la serie de diferencias calculada desde $\mathcal{X}'$, donde $d_i = \mathcal{X}'_i - \mathcal{X}'_{i-1}$.

Ha de señalarse que todo $d_i \in X_D$ es un valor en el intervalo [-1,1], como consecuencia del proceso de normalización.

2.2 Etiquetado

La serie de diferencias representa la evolución de las pendientes a lo largo de la serie. El rango de todas las posibles pendientes se divide en grupos y se asigna una etiqueta cualitativa a cada grupo.

Seguidamente presentamos la colección de grupos y la etiqueta asignada a los rangos de pendientes. Este alfabeto se crea con el parámetro $\delta = 5$, como se presentó en [6], y no se aplican las restricciones presentas en el lenguaje SDL definido en [2].

Tabla 2. En esta tabla, la primera columna representa la etiqueta cualitativa para cada rango de pendientes, que se muestra en la segunda columna. La última columna contiene el carácter asignado a cada etiqueta.

Etiqueta	Rango	Simbolo
Alto incremento	(.2,1]	H
Incremento medio	(.04,.2]	M
Incremento pequeño	(0,.04]	L
Sin variación	0	0
Pequeño decremento	[-.04,0)	l
Decremento medio	[-.2,-.04)	m
Alto decremento	[-1,-.2)	h

Este alfabeto se usa para obtener la cadena de caracteres $S_X = \langle c_1, \mathrm{K}\ , c_{t-1} \rangle$ correspondiente a la serie temporal X. donde cada c_i representa la evolución de la curva entre dos instantes de tiempo adyacentes. Se obtiene de $X_D = \langle d_0, \mathrm{K}\ . d_{t-1} \rangle$ asignando a cada d_i su carácter de acuerdo con la tabla superior.

Esta traducción de la serie temporal en una secuencia de símbolos nos abstrae de los valores reales y nos centra en la forma de la curva. Cada secuencia de símbolos describe una familia completa de curvas con una evolución similar.

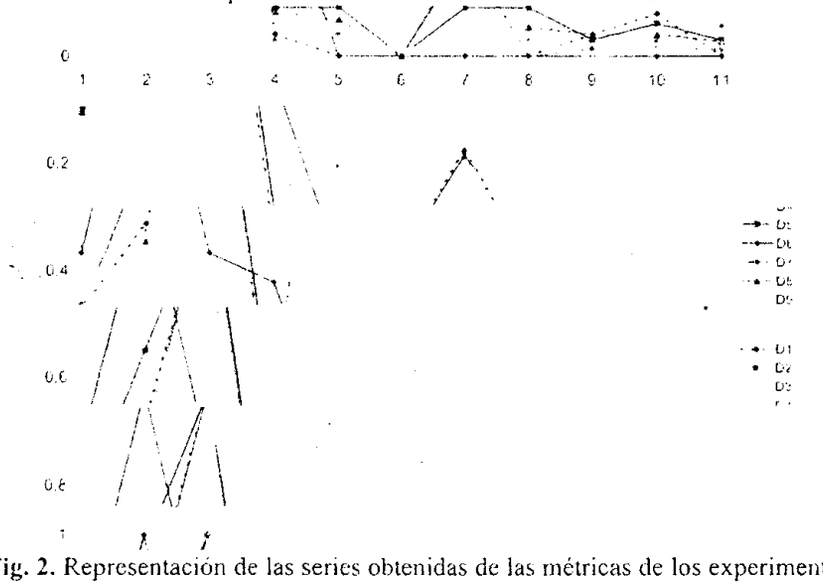

Fig. 2. Representación de las series obtenidas de las métricas de los experimentos.

2.3 Definición de Similitud QSI

Sean S_λ, S_γ las cadenas obtenidas de X, Y normalizadas y etiquetadas. El índice de similitud QSI entre las cadenas S_λ, S_γ se define como

$$QSI(S_\lambda, S_\gamma) = \frac{\nabla(LCS(S_X, S_\gamma))}{m} \qquad (1)$$

donde ∇S es el cuantificador aplicado a la cadena S. El cuantificador proporciona el número de caracteres de S. Además m se define como $m = \max(\nabla S_X, \nabla S_Y)$. Así, la similitud QSI puede entenderse como el número de símbolos que aparecen en el mismo orden en ambas cadenas divido por el tamaño de la cadena más larga.

3. Aplicación del QSI

Para la aplicación del índice QSI a los valores de las métricas de los diferentes experimentos, se considera cada serie de valores como una serie temporal. Las series obtenidas con esta consideración, ya normalizadas, se presentan en la Fig. 2.

Se realizó el cálculo del índice QSI y se obtiene el dendograma que se muestra en la figura 3.

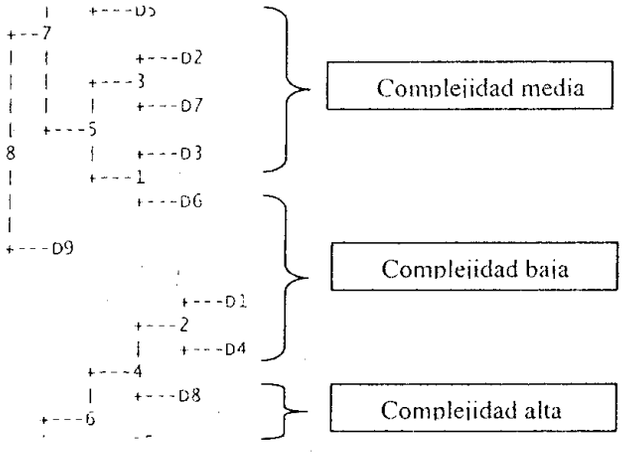

Fig. 3. Dendograma obtenido al aplicar QSI

4. Comparación y análisis de los resultados

Como se puede observar analizando la figura 1 y 3, con el índice QSI se obtiene una agrupación de los diagramas (ver figura 2) diferente a la obtenida anteriormente. Para analizar el significado de la diferencia del agrupamiento obtenido y decidir cual de los resultados es el más representativo decidimos considerar además del valor de las métricas el valor medio del tiempo de mantenimiento para cada diagrama, obtenido en el experimento realizado (ver tablas 3 y 4).

Si nos centramos en la última columna de las tablas 3 y 4 y observamos en que prototipo figura cada diagrama, veremos que considerando los tiempos de mantenimiento parecería ser que el agrupamiento obtenido con el clustering jerárquico usado en los "Repertory Grids" [3] refleja mejor la realidad con respecto al tiempo que puede consumir mantener dada diagrama de clases.

Este resultado debe ser considerado como preliminar, no podemos realizar una afirmación sobre la bondad o no bondad del QSI, creemos que será necesario

aplicarlo a otras muestras de datos (de mayor tamaño) para obtener mayor evidencia de su funcionamiento.

Tabla 3. Agrupamiento de los diagramas obtenidos en [9] incluyendo el tiempo de mantenimiento

	NC	NA	NM	Nassoc	NAgg	NDep	NGen	NaggH	NgenH	MaxDIT	MaxHAgg	Tiempo promedios de mantenimiento (minutos)
Diagramas de complejidad baja												
D1	7	11	22	1	0	0	5	0	1	2	0	2,696
D2	8	12	31	1	6	0	1	1	1	1	2	3,043
D3	3	17	24	2	0	0	0	0	0	0	0	2,913
D6	7	16	7	6	0	0	0	0	0	0	0	2,739
D5	9	19	29	3	3	0	3	3	1	2	1	3,345
Diagramas de media												
D7	23	33	66	4	5	2	16	2	3	3	3	4,87
D8	20	30	65	6	5	0	14	4	3	3	2	4,826
D4	10	12	21	15	3	0	0	2	0	0	1	4,087
Diagramas de complejidad alta												
D9	23	65	80	20	3	2	3	3	1	2	3	7,087

Tabla 4. Agrupamiento de los diagramas utilizando el QSI, incluyendo el tiempo de mantenimiento.

	NC	NA	NM	Nassoc	NAgg	NDep	NGen	NaggH	NGenH	MaxDIT	MaxHAgg	Tiempo promedio de mantenimiento (minutos)
Diagramas de complejidad baja												
D1	7	11	22	1	0	0	5	0	1	2	0	2,696
D4	10	12	21	15	3	0	0	2	0	0	1	4,087
D8	20	30	65	6	5	0	14	4	3	3	2	4,826
D5	9	19	29	3	3	0	3	3	1	2	1	3,345
Diagramas de complejidad media												
D2	8	12	31	1	6	0	1	1	1	1	2	3,043
D7	23	33	66	4	5	2	16	2	3	3	3	4,87
D3	3	17	24	2	0	0	0	0	0	0	0	2,913
D6	7	16	7	6	0	0	0	0	0	0	0	2,739
Diagramas de complejidad alta												
D9	23	65	80	20	3	2	3	3	1	2	3	7,087

5. Trabajo Futuro

Nuestros próximos esfuerzos se centrarán en el estudio de un mecanismo que permita la elección del número de regiones cualitativas en que dividir los rangos y sus puntos de corte de una forma automática. Otros avances futuros para el índice QSI serán el análisis de su comportamiento frente al ruido en los datos y el proporcionar información más precisa sobre las características de las zonas que se detecten como similares para permitir al usuario su ponderación dependiendo del dominio del problema.

Además continuando con el problema abordado en este trabajo, aplicaremos el QSI a datos obtenidos en futuras réplicas del experimento considerado y también a datos relativos a proyectos reales intentando sobretodo conseguir una mayor cantidad de diagramas de clases de distinta complejidad.

Agradecimientos

Este trabajo se ha realizado dentro del proyecto DOLMEN financiado por la Subdirección General de Proyectos de Investigación - Ministerio de Ciencia y Tecnología (TIC 2000-1673-C06-06).

Referencias

1. Agrawal R., Faloutsos C. y Swami A., Efficient similarity search in sequence databases. In Proc. of the Fourth Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '93), Chicago, 1993.
2. Agrawal R., Psaila G., Wimmers E.L. and Zaït M.,Querying shapes of Histories. The 21st VLDB Conference Switzerland, 1995, 502-514.
3. Bell R., Analytic Issues in the Use of Repertory Grid Technique. Advances in Personal Construct Psychology 1, 25-48, 1990.
4. Chan K. and Wai-chee F.A.,Efficient time series matching by wavelets Proc. 15th International Conference on Data Engineering, 1999.
5. Chu S., Keogh E., Hart D., Pazzani M., Iterative Deepening Dynamic Time Warping for Time Series. To appear in the Second SIAM International Conference on Data Mining (SDM-02), 2002 .
6. Cuberos F.J., Ortega J.A., Gasca R.M. and Toro M., QSI - Qualitative Similarity Index. 16th Intl. Workshop of Qualitative Reasoning, Sitges. Spain, 2002, pp ¿?¿?¿
7. Cormode G., Muthukrishnan S., Paterson M., Sahinalp S.C. and Vishkin U. , Techniques and applications for approximating strong distances - Rough Draft. http://citeseer.nj.nec.com/320221.html , 2001.
8. Faloutsos C., Ranganathan M., and Manolopoulos Y., Fast subsequence matching in time-series databases. The ACM SIGMOD Conference on Management of Data, 1994, 419-429.
9. Genero M., Olivas J., Romero F., Piattini M. Assesing the maintnability of OO conceptuals models. First International Workshop on Conceptual Modeling Quality (IWCMQ'02). LNCS (to appear), 2002.

10. Goldin D.Q. and Kanellakis P.C., On similarity queries for time-series data: constraint specification and implementation. In 1st Intl. Conf. on the Principles and Practice of Constraint Prog., Minneapolis, 1994, 419-429.
11. Huhtala Y., Kärkkäinen J. and Toivonen H., Mining for similarities in aligned time series using wavelets. Data Mining and Knowledge Discovery: Theory, Tools, and Technology. SPIE Proc. Vol. 3695.
12. Kahveci T. and Singh A. , Variable length queries for time series data. In Proceedings of the 17th Intl. Conf. on Data Engineering, Heidelberg, 2001.
13. Kahveci T., Singh A. and Gurel A.,Similarity Searching for Multi-Attribute Sequences, SSDBM 2002, Edinburgh, Scotland, 2002.
14. Keogh E.J. and Smyth P., A probabilistic approach to fast pattern matching in time series databases, Proceedings of the 9th International Conference on Tools with Artifial Inteligence, IEEE Press, 1998, 578-584.
15. Keogh E.J. and Pazzani M.J.,An enhaced representation of time series wich allows fast and accurate classification, clustering and relevance feedback, Proc. 4th International Conference of Knowledge Discovery and Data Mining, AAAI Press, 1998,pp. 239-241.
16. Keogh E.J. and Pazzani M.J.,A simple dimensionality reduction technique for fast similarity search in large time series databases, In Proc. of the Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000.
17. Keogh E. And Pazzani M., Scaling up Dynamic Time Warping for Datamining Applicactions. Proc. of the 6th ACM SIGKDD International Conf. On Knowledge Discovery and Data Mining.Boston,USA., 2000, 285-289.
18. Keogh E.J., Chakrabarti K., Mehrotra S. and Pazzani M.J.,Locally adaptative dimensionality reduction for indexing large time series databases,, In Proc. of ACMSIGMOD, 2001, 151-162.
19. Keogh E.J., Chakrabarti K., Pazzani M.J. and Mehrotra S., Dimensionality Reduction for Fast Similarity Search in Large Time Series Databases, Knowledge and Information Systems Vol 3, 263-286, 2001.
20. Kim S-W, Park S. and Chu W.W., An Index-Based Approach for Similarity Search Supporting Time Warping in Large Sequence Databases. In Proc. 17th IEEE Int'l Conf. on Data Engineering, Heidelberg, Germany, 2001.
21. Park S., Lee D., and Chu. W. W., Fast retrieval of similar subsequences in long sequence databases. In Proc. 3rd IEEE Knowledge and Data Engineering Exchange Workshop (KDEX), Evanston, IL, 1999, 60-67 .
22. Park S., Lee D., and Chu. W. W., Segment-based approach for subsequence searches in sequence databases.Proceedings of the 2001 ACM Symposium on Applied Computing (SAC), March 11-14, Las Vegas, NV, USA, 2001.
23. Rafiei D. and Mendelzon A., Similarity-based queries for time data series. In Proc. of the ACM SIGMOD Intl. Conf. of Management of Data(SIGMOD '97), Tucson, 1998, 13-24.
24. Rafiei D. and Mendelzon A., Efficient Retreival of similar time sequences using DFT. In Proc. of the 5th Intl. Conf. on Foundations of Data Organization and Algorithms (FODO '98). Kobe, 1998.
25. Shatkay H. and Zdonic S.,Approximate queries and representation for large data sequences.In Proc. of the 12th International Conference on Data Engineering, 1996, 546-553.
26. Yi B.K., Jagadish H. and Faloutsos C., Efficient retrieval of similar time sequences under time warping. IEEE Intl. Conf. on Data Engineering, 1998, 201-208.
27. Yi B.K. and Faloutsos C., Fast time sequence indexing for arbitrary L_p norms. Proceedings of the 26th Intl. Conf. on Very Large Databases, Cairo, 2000.