IX Workshop de Investigadores en Ciencias de la Computación (2007)
Pág. 487-491

PROCESO DE AGREGACIÓN CON MÚLTIPLES NIVELES DE EVIDENCIA PARA ESTUDIOS EXPERIMENTALES EN INFORMÁTICA

Enrique Fernández[1,2], Paola Britos[1,2], Oscar Dieste[3], Ramón García-Martínez[1,2]

[1] Centro de Ingeniería de Software e Ingeniería del Conocimiento. Escuela de Postgrado. ITBA. Argentina
[2] Laboratorio de Sistemas Inteligentes. Facultad de Ingeniería. Universidad de Buenos Aires. Argentina
[3] Facultad de Informática. Universidad Politécnica de Madrid. España.
rgm@itba.edu.ar

RESUMEN

En el presente trabajo se describe el estado de avance de la línea de investigación sobre estrategias de agregación de estudios experimentales en Informática con énfasis en el desarrollo de métodos que permitan paliar las falencias de los estudios documentados.

1. Introducción

1.1. Meta-Análisis

Según [Cochrane, 2007] el meta-análisis es el análisis estadístico de una serie de estudios individuales, con el objeto de integrar los resultados en una medida resumen. El propósito del meta-análisis es combinar resultados numéricos. Para lo cual, a partir de los estimadores de efecto de los distintos estudios se calcula un estimador de efecto global. El meta-análisis es conceptualmente muy sencillo: el estimador de efecto global se calcula como una media ponderada de los estimadores de efecto de los estudios. Por ejemplo, las diferencias de riesgo individuales de los estudios se ponderan y se calcula la media; el resultado es la diferencia de riesgo combinada que tomamos como resultado del meta-análisis. Al realizar un meta-análisis, deseamos hallar un resultado numérico que sea resumen representativo de los resultados de los estudios individuales, y por tanto que signifique una mejora sobre las estimaciones individuales. Idealmente, el meta-análisis debe partir de los estudios individuales -con sus virtudes y defectos- y obtener un resultado que sea más fiable que los resultados individuales de los que partíamos.

1.2. Técnicas de Agregación

1.2.1. Effect Size

La técnica Effect Size permite estimar el efecto o mejora de un determinado tratamiento respecto de otro. En la actualidad existen varias adaptaciones de la formula de estimación, las mismas se diferencian básicamente en la cantidad de sujetos experimentales tratados en los estudios experimentales, es decir si la cantidad de sujetos experimentales es alta o no y si los distintos estudios experimentales seleccionados tienen la misma cantidad de sujetos.
Dentro de este contexto, a continuación se describe la variante de la técnica que permite trabajar con estudios de pocos sujetos experimentales (es difícil conseguir dentro del ámbito de la Informática estudios con gran cantidad de sujetos experimentales) y permite combinar estudios con cantidades de sujetos experimentales diferentes.

1.2.1.1. Estimación de Efectos para Estudios Individuales

A continuación, se describe la formula corregida de [Hedges y Olkin, 1985] , la misma es la utilizada y recomendada en los procesos de meta-análisis de [Cochrane, 2007];

$$d = J(N-2)\frac{\mu^E - \mu^C}{\sigma_P}$$

d =	Tamaño del efecto
μ^E=	Media del tratamiento nuevo (o tratamiento principal de la comparación)
μ^C =	Media del tratamiento de contraste (o tratamiento secundario)
σ_p =	Desvío estándar conjunto de los tratamientos
J (N – 2) =	Es un factor de ajuste tabulado
N =	Es la cantidad total de sujetos experimentales ($n_t + nc_c$)

1.2.1.2. Estimación de Efectos para una Serie de Estudios

La forma básica de calcular el tamaño del efecto para una serie de experimentos, consiste en sumarizar los tamaños de efectos particulares ponderados por un coeficiente de peso estimado en función de la cantidad de sujetos experimentales [Hedges y Olkin, 1985]:

$$dw = w_1 * d_1 + ... + w_k * d_k$$

dw = es el tamaño del efecto global o total
$w_{1 \ldots k}$ = son los pesos individuales del experimentos
$d_{1 \ldots k}$ = son los tamaños de efectos de los experimentos

La formula para calcular los pesos es la siguiente:

$$w_i = \frac{\tilde{n}_i}{\sum_{j=1}^{k} \tilde{n}_j}$$

$\tilde{n}_i$ = es $n_i^e * n_i^c / (n_i^e + n_i^c)$
n_i^e = Cantidad de sujetos experimentales del tratamiento nuevo (o técnica principal de la comparación)
n_i^c = Cantidad de sujetos experimentales del tratamiento de contraste (o técnica secundaria)

Estimación del intervalo de confianza con un nivel de exactitud del 95 %:

$$\delta_L = d_w - 1.96\sqrt{v} \qquad \delta_L = d_w + 1.96\sqrt{v}$$

δ_L = cota del intervalo de confianza inferior.
δ_U = cota del intervalo de confianza superior.
dw = es el tamaño del efecto global o total
v = representa la varianza estimada por formula 3-10

1.2.2. Vote Counting

Vote Counting [Hedges y Olkin, 1985] es una técnica que, a través de un conjunto acotado de variables estadísticas (signo de la deferencia entre las medias de tratamiento y cantidad de sujetos experimentales), permite inferir un tamaño de efecto para un conjunto de estudios experimentales dado. Esto se hace mediante un proceso de inferencia iterativo, en el cual, en primer lugar se genera una lista de valores de efecto. Luego de esto se intenta asignar un valor de probabilidad para cada uno de los efecto en base a un parámetro obtenido mediante la combinación del signo de la diferencia entre medias y la cantidad de sujetos experimentales. A continuación se describe la formula a aplicar para obtener el tamaño de efecto aplicando la técnica de conteo de votos:

$$L(\delta \mid X_1, ..., X_i) = \sum_{i=1}^{k} \left\{ X_i \ln\left[1 - \phi\left(-\sqrt{\tilde{n}}\delta\right)\right] + (1 - X_i) \ln \phi\left(-\sqrt{\tilde{n}}\delta\right) \right\}$$

L = representa la lista de probabilidad para cada tamaño de efecto, de la cual se deberá seleccionar el efecto de mayor probabilidad.
∂ = es la media de los efecto, para lo cual se debe generar, en primera instancia, un conjunto de valores posibles, para luego mediante un proceso iterativo se deberá seleccionar el de mayor probabilidad.
X_i = representa a la diferencia entre las medias, y tomará el valor 1 si la media del tratamiento principal es mayor a la de contraste y 0 cuando es menor o igual.
$\tilde{n}_i$ = es $n_i^e * n_i^c / (n_i^e + n_i^c)$
k = representa a la cantidad de estudios experimentales.

Estimación de la Varianza:

$$Var = \left(\sum_{i=1}^{k} D_i \right)^{-1}$$

Var = representa a la varianza del tamaño de efecto.
D_i = representa al factor estimado en la formula 3-9
k = representa a la cantidad de estudios experimentales.

Estimación del parámetro D:

$$D_i = \sqrt{\frac{\tilde{n}_i}{2\Pi}} \left(-\frac{1}{2} \tilde{n} \delta^2 \right)$$

D_i = representa un factor de relación que debe estimarse para cada uno de los estudios experimentales.
$\tilde{n}_i$ = es $n_i^e * n_i^c / (n_i^e + n_i^c)$
∂ = es la media de los efecto estimada.

Estimación del intervalo de confianza con un nivel de exactitud del 95 %:

$$\delta_L = \delta - 1.96\sqrt{\text{var}} \qquad \delta_U = \delta + 1.96\sqrt{\text{var}}$$

δ_L = cota del intervalo de confianza inferior. δ_U = representa la cota del intervalo de confianza superior.

δ = representa al tamaño de efecto estimado *Var* = representa la varianza estimada

1.2.3. Conteo De Votos Directo

El conteo de votos directo, es una técnica que tiene como objetivo indicar que tratamiento tiene mayor cantidad de estudios experimentales a su favor. No requiere de publicaciones de variables estadísticas especiales, solo es necesario saber si alguno de los tratamientos tuvo mejores resultados que el otro. Para ello se debe sumarizar los resultados de los estudios dentro de tres categoría: a favor del tratamiento principal (la media del tratamiento principal es mayor a la media del tratamiento de contraste), a favor del tratamiento de contraste (la media del tratamiento principal es menor a la media del tratamiento de contraste) y neutro (no existe diferencias entre las medias de ambos tratamiento). Una vez sumarizados los resultados se deberá analizar cual de las categoría obtuvo la mayor cantidad de votos, en base a esto se podrá establecer que el tratamiento principal permite obtener resultados mejores, iguales o peores que el tratamiento de contraste.

2. Estado de la cuestión. Experimentación en la Informática

En Informática, de acuerdo a la norma 610.12 de IEEE, se debe aplicar conocimiento científico para el desarrollo, operación y mantenimiento de sistemas software. Para ello cuenta con métodos, técnicas y herramientas para ser utilizadas en cada actividad de acuerdo a las condiciones que se disponga. Sin embargo, en la actualidad generalmente no se cuenta con técnicas ni métodos que cuenten con una justificación científica ni un "estudio objetivo de su efectividad" [Juristo y Moreno, 2001]. Por lo tanto, es necesario un marco que permita a los ingenieros poder conocer cuales son los mejores métodos y herramientas que se deben aplicar a través de un método científico y por lo tanto objetivo. Este marco es la investigación experimental, utilizada también en otros ámbitos para brindar información objetiva sobre hipótesis que se desean probar. De esta forma, como afirma Pfleeger [1999], se "permitirá ganar más entendimiento de que hace un software bueno y como hacerlo mejor".

3. Definición del problema

Si bien la experimentación dentro del ámbito de la Informática ha ido creciendo y mejorando en los últimos años, a la fecha la mayoría de los estudios empíricos poseen falencia bastante graves. Entre las falencias mas importantes podemos señalar:

- Estudios empíricos hechos con pocos sujetos experimentales [Burton *et al*., 1990].
- Estudios empíricos con sesgos de publicación [Zmud *et al*., 1993].
- Falta de estandarización en variables respuesta en estudios que a priori analizan el mismo fenómeno [Schweickert *et al*.,1987] y [Burton *et al*., 1990].
- Falta de standardización en la forma de referenciar a los tratamientos [Agarwal *et al*., 1990] y [Woody *et al*., 1996].
- Estiman factores estadísticos que requieren el conocimiento de la distribución, cuando esta no se puede determinar por la baja cantidad fe sujetos experimentales [Burton *et al*., 1988].
- Falta de verdaderas replicaciones de estudios
- Muchos de los estudios empíricos son de baja calidad [Crandall, 1989].

Estos aspectos hacen que la aplicación de un proceso de agregación estándar [Kitchenham, 2004], mediante el cálculo de tamaño de efectos, sea prácticamente imposible [Davis *et al*., 2006] y [Brereton *et al*., 2004].

4. Solución propuesta

Para solucionar el problema de agregación de estudios, se propone una estrategia de agregación **multinivel**, la misma provee diferentes niveles de evidencia en función de la calidad de los estudios experimentales seleccionados, tanto a nivel de realización del estudio como a nivel de la calidad del reporte presentado (fundamentalmente variables estadísticas publicadas). Cada uno de estos niveles de evidencia esta asociado a una técnica de agregación específica.

4.1. Niveles de Evidencia

Para identificar los niveles de evidencia se usaran tres términos que a menudo se utilizan dentro del ámbito del derecho. Estos términos son: Evidencia, Presunción y Sospecha, y su significado según el diccionario es:

- Evidencia: hecho que no deja lugar a dudas sobre su veracidad.
- Presunción: hecho que se considera verdadero hasta que se demuestre lo contrario.
- Sospecha: suponer un hecho por conjeturas basadas en apariencia.

En tal sentido, a continuación se describe que tipo de estudio será asignado a cada nivel de evidencia para obtener la calidad de respuesta deseada:

- ***Evidencia,*** para este nivel de agregación se seleccionarán los mejores estudios experimentales encontrados, en cuanto a la confección de los mismos y de los datos descriptos en el reporte publicado.
- ***Presunciones***, para este nivel de agregación se seleccionarán todos los estudios experimentales del nivel anterior mas los estudios con sesgo de publicación leve.
- ***Sospechas***, para este nivel de agregación se seleccionarán todos los estudios experimentales de los nivel anterior mas los estudios con sesgo de publicación medio o grave.

El objetivo de este proceso es permitir incorporar al proceso de agregación la mayor cantidad de estudios experimentales posibles y generando conclusiones de diversos niveles. A continuación, se describe la relación existente entre las técnicas de agregación y los niveles de calidad de los estudios:

Técnicas de agregación	**Motivo de la asignación de estudios**
Effect Size	Para esta técnica solo se aceptan estudios que no posean sesgos y sean similares en cuanto a su confección y dominio de aplicación. Esto se debe a que, por un lado se requiere que el estudio no posea sesgos de publicación y por otro una alta calidad en el desarrollo del mismo.
Vote Counting	Para esta técnica se podrán aceptar todos los estudios del nivel anterior mas los estudios con sesgos leves de publicación, por que la misma no necesita conocer cuan superior es un tratamiento respecto del otro ni las varianzas.
Conteo de votos directo	Para esta técnica podrán aceptarse todos los estudios de los niveles anteriores, mas los estudios con sesgos graves de publicación, ya que esta técnica la única restricción de aplicación que tiene es saber si alguno de los tratamientos a dado mejores resultados que el otro.

4.2. Interpretación de los resultados

Los resultados obtenidos deben analizarse desde los más confiables a los menos confiables. Es decir en primer lugar se debe analizar los resultados obtenidos mediante Effect Size, luego los obtenidos por Vote Counting y por último los obtenidos por Conteo de Votos Directo. Cuando todas las técnicas arrojen resultados compatibles (den que el tratamiento x es mejor que el y) podrá decirse que toda la evidencia obtenida corrobora la misma hipótesis, pero, si esto no es así se deberá hacer un análisis mas detallado de los estudios menos fiables para intentar determina si es que existen

variables independientes no identificas que están modificando los resultados obtenidos, un proceso similar al que se realiza cuando se lleva a delante un análisis de Homogeneidad y Sensibilidad.

5. Conclusiones

El definir una estrategia de agregación de más de un nivel evita la generación de conclusiones a criterio personal de quien realiza el proceso de revisión. Esto se debe a que cuando solo se intenta aplicar la técnica de Effect Size en la mayoría de los casos el proceso de agregación (dentro del ámbito de la Informática) queda desierto.

6. Futuras líneas de investigación

Incorporar una estrategia para verificar los factores de "Homogeneidad" (que los estudios incluidos en el meta-análisis sean realmente compatibles) y "Sensibilidad" (comprobar la influencia de los estudios individuales en la estimación del efecto), para garantizar de esta manera la fiabilidad de las conclusiones generadas.

7. Formación de recursos humanos

En la línea de investigación cuyos resultados parciales se reportan en esta comunicación, se encuentran trabajando: un tesista de doctorado en informática, un tesista de maestría en ingeniería del software y tres investigadores formados (uno español y dos argentinos).

8. Bibliografía

Agarwal, R.; Tanniru, M.; 1990; *Knowledge Acquisition Using Structured Interviewing: An Empirical Investigation;* Journal of Management Information System, 7(1).

Brereton, P.; Kitchenham, B.; Budgen, D.; Turner, M.; Khelil, M.; 2004; *Employing Systematic Literature Review: An Experimental Report*.

Burton, A., Shadbolt, N., Hedgecock, A. y Rugg, G. 1988. *A Formal Evaluation of Knowledge Elicitation Techniques for Expert Systems: Domain 1*. Proceedings of Expert Systems '87 on Research and Development in Expert Systems IV**.** Pág. 136-145.

Burton, A., Shadbolt, N., Rugg, G. y Hedgecock, A. 1990. *The Efficacy of Knowledge Elicitation Techniques: A Comparison Across Domains and Level of Expertise*. Knowledge Acquisition 2(2): 167-178.

Cochrane; 2007; *Curso Avanzado de Revisiones Sistemáticas*; www.cochrane.es/?q=es/node/198. Página vigente al 21/03/07.

Crandall Klein, B. y Asociados; 1989. *A Comparative Study Of Think-Aloud And Critical Decision Knowledge Elicitation Method*. SIGAR Newsletter, 108: 144-146.

Davis, A.; Dieste o.; Hickey, A.; Juristo, N.; Moreno, A.; 2006; *Effectiveness of Requirements Elicitation Techniques: Empirical Results Derived from a Systematic Review*; 14th IEEE International Requirements Engineering Conference (RE'06) pp. 179-188

Hedges, L.; Olkin, I.; 1985; *Statistical methods for meta-analysis*. Academic Press.

Juristo N. y Moreno A. 2001. *Basics of Software Engineering Experimentation*. Kluwer

Kitchenham, B. A. 2004; Procedures for performing systematic reviews. Keele University; TR/SE-0401. Keele University Technical Report.

Schweickert, R., Burton, A., Taylor, N., Corlett, E., Shadbolt, N., Rugg, G. y Hedgecock, A.; 1987. *Comparing Knowledge Elicitation Techniques: A Case Study*. Artificial Intelligence Review (1): 245-253.

Woody, J.; Will, R.; Blanton, J.; 1996; *Enhancing Knowledge Elicitation using the Cognitive Interview;* Expert system with application; Vol. 10 N. 1

Zmud, R.; Anthony, W.; Stair R.; 1993; *The Use of Mental Imagery to Facilitate Information Identification in Requirements Analysis;* Journal of Management Inf. System; 9(4).