Hacia una metodología para la selección de técnicas de depuración de datos

Towards a methodology for selection of data cleansing techniques

Iván Amón Uribe Esp., Claudia Jiménez Ramírez Ph.D.
Grupo de Investigación y Desarrollo en Inteligencia Artificial-Universidad Nacional de Colombia
{iamonu, csjimene}@unalmed.edu.co

Recibido para revisión 26 de enero de 2009, aceptado 20 de mayo de 2009, versión final 9 de junio de 2009

Resumen-Errores de digitación, datos inconsistentes, valores ausentes o duplicados, son algunos de los problemas que pueden presentar los datos almacenados en las bases y bodegas de datos, deteriorando su calidad y en consecuencia, la calidad de las decisiones que se tomen con base en el nuevo conocimiento obtenido a partir de ellos. Este artículo pone de manifiesto la necesidad de una guía metodológica que apoye a los analistas de datos en la selección de las técnicas de depuración, considerando los diferentes tipos de errores en los datos y la naturaleza de los mismos.

Palabras Clave-Bases y Bodegas de Datos, Minería de Datos, Limpieza de datos, Preprocesamiento de datos, Calidad de datos.

Abstract-Typing errors, inconsistent data, missing values or duplicates, are some of the problems that can be present in databases and data warehouses, affecting data quality and thus the quality of decision making based on the new knowledge extracted from them. This article highlights the need for a methodological support to data analysts in the selection of cleansing techniques, considering different types of data errors and their nature.

Keywords-Databases, Data Warehousing, Data Mining, Data Cleansing, Data Preprocessing, Data Quality.

I. INTRODUCCIÓN

Actualmente, las organizaciones toman decisiones, cada vez más, con base en el conocimiento derivado de los datos almacenados en sus bases o bodegas de datos, aplicando el enfoque denominado Inteligencia del Negocio (Business Intelligence, en inglés). Por tanto, es de vital importancia que los datos contengan la menor cantidad de errores posibles.

Dasu et al.[1], en el año 2003, afirmaban que "es bastante común que las bases de datos tengan del 60% al 90% de problemas de calidad en los datos".

En el mismo sentido, una investigación realizada por la firma Gartner, en el año 2007 [2], indica que más del 25% de los datos críticos en las compañías presentan errores. Según Andreas Bitterer, vicepresidente de investigación de Gartner, "No existe una compañía en el planeta que no tenga un problema de calidad en sus datos y aquellas compañías que reconocen tenerlo, a menudo subestiman el tamaño de éste". Aunque idealmente, los datos almacenados no deberían contener errores, es casi inevitable que existan y merecen toda la atención para poder hacer inferencias válidas y la toma de decisiones acertadas, aplicando el nuevo enfoque gerencial BI.

Los datos "sucios" pueden conducir a decisiones erróneas, ocasionando pérdidas de tiempo, dinero y credibilidad. Gartner [2] afirma: "la mala calidad de los datos sobre los clientes, lleva a costos importantes, como el sobreestimar el volumen de ventas, e exceso de gastos en los procesos de contacto con los clientes y a la pérdida de oportunidades de ventas. Sin embargo, las empresas están descubriendo que la calidad de sus datos tiene una incidencia significativa en la mayoría de sus iniciativas empresariales estratégicas, y no sólo en el área de ventas o investigación de mercados. Otras funciones como elaboración de presupuestos, producción y distribución también se ven afectadas".

La importancia de contar con datos confiables, con los cuales se puedan tomar decisiones acertadas, es cada vez mayor. Conceptos como Gestión del Conocimiento, Minería de Datos e Inteligencia de Negocios, se están desarrollando a pasos agigantados, y de poco o nada sirven si se basan en datos errados.

Para ilustrar los posibles errores en los datos y los problemas que originan, se puede tomar como ejemplo la base de datos ScienTI, donde se registra la actividad investigativa de Colombia por parte de Colciencias. En ella, cada investigador actualiza sus datos por medio del programa CvLAC (Curriculum vitae Latinoamericano y el Caribe) registrando los proyectos de investigación en los que participa. Dado que en un proyecto (i) pueden participar varios investigadores, (ii) cada uno de ellos ingresa sus datos al sistema por separado y (iii) no existe una identificación única de los proyectos, puede suceder que el nombre de un mismo proyecto no se escriba exactamente igual por parte de todos sus integrantes (por ejemplo: "Metodología para selección de técnicas para depuración de datos" vs "Metodología para selección de técnicas de depuración de datos"). Si no se toman las medidas adecuadas, previas a la contabilización de la cantidad de proyectos, se obtendrá un número que sobredimensiona la producción académica de las universidades o centros y se distorsionará la realidad.

Para los distintos errores que pueden presentar los datos, diversos investigadores han desarrollado técnicas para detectarlos y corregirlos. Los trabajos relacionados con esta problemática se agrupan bajo diferentes denominaciones como calidad de datos (Data Quality), heterogeneidad de datos (Data Heterogeneity), limpieza de datos (Data Cleansing) o reconciliación de datos (Data Reconciliation),

Dada la multiplicidad de técnicas existentes para la depuración de los datos, no es trivial decidir cuales deben ser usadas en cada caso particular. En este artículo se consignan los primeros pasos hacia la construcción de una guía metodológica que apoye a los analistas de datos en la selección de las técnicas a ser aplicadas a un conjunto de datos en particular, de acuerdo con la naturaleza y la distribución de los datos analizados.

El resto del presente artículo está estructurado como sigue. La Sección 2, presenta algunos trabajos realizados sobre calidad de datos. La Sección 3, plantea la necesidad de una metodología para la selección de técnicas de depuración de datos. La Sección 4, bosqueja la aproximación metodológica para seleccionar la técnica de depuración y por último se presentan las conclusiones y trabajos futuros.

II.TRABAJOS RELACIONADOS

Múltiples trabajos se han realizado en la temática de calidad de datos. A continuación, se relacionan algunos que son de interés para el propósito de este artículo: trazar un camino que conduzca a lograr una metodología para seleccionar las técnicas apropiadas para aplicar a una determinada tarea de depuración de datos.

En cuanto a trabajos relacionados con la clasificación y la detección de los problemas, son varios los que han realizado clasificaciones de las anomalías en los datos [3, 4, 5], pero Oliveira et. al. [6] no sólo realizan una taxonomía con treinta y cinco problemas de calidad de los datos, sino que plantean métodos semiautomáticos para detectarlos, los cuales representan mediante árboles binarios. Los árboles corresponden al razonamiento que se necesita hacer para detectar un problema particular.

Un tipo de problema que pueden presentar los datos es el conocido como Record Linkage o detección de duplicados [7], que tiene como meta identificar registros o tuplas que se refieran a la misma entidad del mundo real, aún si los datos no son idénticos. Esto es, se trata de la detección de atributos o registros que tienen contenidos distintos pero que debieran ser el mismo. En 1946, H. L. Dunn [8], de la oficina del censo de los Estados Unidos, introdujo el término en esta forma: "Cada persona en el mundo crea un Libro de Vida. Este libro comienza con su nacimiento y termina con su muerte. Record Linkage es el nombre del proceso de ensamblar las páginas de ese libro en un volumen".

El Record Linkage computarizado fue planteado primero por el genetista Canadiense Howard Newcombe y sus colaboradores en 1959 y continuó sus publicaciones sobre el tema a lo largo de treinta años [9-13]. Otro autor reconocido es William Winkler, autor prolífico quien ha publicado múltiples artículos sobre el tema [14-30].

Uno de los trabajos más recientes es el resumen de Elmagarmid et. al. [31] en el cual se exponen las principales técnicas para el problema de la detección de duplicados, tanto para registros completos como para campos tipo texto en forma individual.

Otro tipo de problemas es el de los valores extremos atípicos, conocidos como Outliers [32]. Aunque no necesariamente son errores, pueden ser generados por un mecanismo diferente de los datos normales como problemas en los sensores, distorsiones en el proceso, mala calibración de instrumentos y/ o errores humanos. También sobre este tema se han realizado múltiples investigaciones, entre las cuales se encuentran trabajos tipo resumen [33], trabajos comparativos [34-35], trabajos sobre técnicas específicas [36-37], entre muchos otros.

En los trabajos realizados, también se encuentran algunos de tipo metodológico. Tierstein [38] presenta una metodología que incorpora dos tareas que se interrelacionan y se traslapan: limpiar los datos de un sistema legacy y convertirlos a una base de datos. En [39], se extienden los sistemas de bases de datos para manejar anotaciones de calidad de los datos en las bases de datos mediante metadatos. En [40], Rittman presenta la metodología seguida por el módulo de Oracle encargado de realizar depuración a los datos (Oracle Warehouse Builder), para realizar este proceso.

Las técnicas desarrolladas por los investigadores hasta el momento, son variadas y casi siempre aplican a un tipo de problema en particular. Es así como existen técnicas para tratar el problema de la detección de duplicados, para detección y

corrección de valores atípicos, para tratar con los valores faltantes y para cada posible problema que puedan presentar los datos.

Para la detección de duplicados, se encuentran técnicas como la distancia de edición [41], distancia de brecha afín [42], distancia de Smith-Waterman [43], distancia de Jaro [44], q-grams [45], Whirl [46] y técnicas fonéticas como soundex [47,48], NYSIIS [49], ONCA [50], Metaphone [51] y Double Metaphone [52].

Para la detección de valores atípicos se encuentran técnicas como el análisis de extremos de Mahalanobis [53], el Método de factor local de extremos (LOF) [54], los modelos basados en reglas [55], la prueba de Tukey y los métodos basados en distancia [56, 57]. Para dar solución a los datos faltantes, existen técnicas como las imputaciones de media, mediana y moda, el algoritmo EM (Expectation-Maximization) [58] y el método de aprendizaje adaptativo, entre otros [59, 60].

III. JUSTIFICACIÓN DE LA METODOLOGÍA

Elmagarmid et al. [31] plantean que ninguna métrica es adaptable a todos los conjuntos de datos. La tarea de depuración de datos, es altamente dependiente de los datos y no está claro si algún día existirá una técnica que domine a todas las demás, en todos los conjuntos de datos.

Lo anterior significa que la calidad de la limpieza lograda sobre los datos, depende de la técnica aplicada y la elección de la técnica está íntimamente ligada con la naturaleza de los datos específicos sobre los que se está trabajando.

La selección de las técnicas para depuración de datos, que permitan la entrega de información confiable para la toma de decisiones, requiere de conocimiento profundo de las propiedades de cada una de las técnicas, sus características y cuándo pueden ser aplicadas con éxito a un conjunto de datos dependiendo de la naturaleza de los mismos.

Para mostrar como la técnica depende del conjunto de observaciones que se deben depurar, se toma como ejemplo la detección de valores extremos. Ésta puede ser fácil con el apoyo de una gráfica que muestre la dispersión de los puntos, pero si se quiere realizar automáticamente la búsqueda y almacenamiento de estos valores, por medio de un procedimiento o una función almacenada, se necesita alguna técnica matemática para hallarlos. Comúnmente se usa la fórmula de Tukey, basada en los cuartiles de la distribución o los valores que subdividen el conjunto de datos ordenados en cuatro partes, cada una con el mismo porcentaje de datos. Tomando como referencia la diferencia entre el primer cuartil Q1 y el tercer cuartil Q3, o el rango intercuartil, se considera un valor extremo o atípico aquel que se encuentra a 1,5 veces esa distancia de uno de esos cuartiles (atípico leve) o a 3 veces esa distancia (atípico extremo). Sin embargo, dependiendo de la distribución de los datos, este método puede fallar. Si el rango intercuartil resulta ser cero, cualquier valor diferente de cero se tomaría como atípico. Por lo tanto, en estos casos, es recomendable usar otro método.

De otra parte, las herramientas comerciales que realizan depuración a los datos, en general, no realizan autónoma y automáticamente este trabajo, sino que requieren la intervención del usuario. Generalmente, ofrecen un conjunto de opciones entre las cuales se debe elegir la técnica a ser aplicada a los datos, tarea que demanda altos conocimientos técnicos.

Teniendo en mente todo lo anterior, surge entonces la pregunta ¿Cómo determinar las técnicas que deben ser empleadas para realizar procesos de depuración a los datos en un caso particular? En la literatura revisada, esta pregunta no se responde satisfactoriamente, ya que no se encontró evidencia de una metodología que indique claramente el procedimiento para seleccionar la técnica más apropiada -bajo alguna(s) métrica(s) predefinidas- a aplicar en una situación específica considerando la naturaleza de los datos en cuestión y el tipo de falla o error que presenten los datos.

Los trabajos de investigación mencionados en la Sección 2, incluyendo aquellos de tipo metodológico, no se ocupan lo suficiente de la selección de las técnicas para depurar los datos, en un caso particular. El trabajo de Oliveira et. al. [6], plantea sin mayor detalle, como detectar la anomalía de los datos sin indicar cual técnica usar para su detección y/o corrección. Tierstein [38], aunque presenta una metodología que intenta cubrir todo el proceso de depuración de datos, se enfoca principalmente hacia el manejo de los datos históricos, no examina las técnicas existentes para depuración, el paso de detección de los defectos no se ocupa de recomendar una técnica y no tiene en cuenta la naturaleza de los datos. Rosenthal et. al. [39], están orientados al enriquecimiento de los sistemas de bases de datos con metadatos, sin examinar ni recomendar técnicas de depuración.

Una metodología ampliamente conocida y usada en proyectos de descubrimiento de conocimiento en bases de datos (KDD: Knowledge Discovery in Databases, en inglés) como Crisp-Dm [61], aunque en su fase de preparación de los datos se ocupa de la transformación y limpieza de los datos, no desciende hasta el nivel de recomendar técnicas específicas dependiendo de la naturaleza de los datos. Similar situación sucede con SEMMA [62], otra metodología de KDD estrechamente ligada a los productos SAS. La metodología seguida por Oracle en [40], confirma que el software ofrecido para la depuración de los datos no selecciona por el usuario la técnica a aplicar. Por lo tanto, la elección de la técnica es fundamental, pero no se conoce de alguna metodología que detalle la forma de realizar esta tarea. Es por esto que este artículo pretende sentar las bases para la construcción de una metodología que oriente al analista de los datos hacia una selección, con mayor rigor científico, de las técnicas para aplicar a un conjunto de datos particular de un dominio específico.

IV. APROXIMACIÓN METODOLÓGICA

A continuación se delinean los principales elementos de lo que será la aproximación o guía metodológica a proponer, tomando inicialmente como base el problema de la detección de duplicados, en los valores de un atributo individual.

Para poder recomendar una técnica a aplicar en un caso específico, es necesario identificar el tipo de problema de los datos, en este caso, el problema de valores que pueden ser diferentes pero que debieran ser el mismo y además conocer como es la distribución de los datos.

La selección de las técnicas hace que se deban conocer en profundidad. Esto es, identificar sus fortalezas y debilidades. Debe establecerse una serie de criterios de evaluación que permitan comparar las técnicas y elegir una de ellas. Los criterios determinarán la eficacia de la técnica ante diferentes situaciones.

Se trata de calificar a cada técnica, según sea eficaz o no, al ser aplicada. La eficacia de una técnica será baja para un criterio, si la similitud entre dos textos decae al presentarse la situación cuando se comparan valores nominales.

Los criterios de evaluación propuestos para detectar duplicados en los textos son:

- **Palabras en orden**: si los dos textos tienen las mismas palabras en el mismo orden o no ('Iván Amón' vs 'Amón Iván').
- **Mayúsculas/minúsculas**: si los textos son muy similares pero con letras mayúsculas o minúsculas indistintamente ('CASA ' vs' casa').
- **Espacios en blanco**: si un texto presenta mayores espacios en blanco que otro, así esté compuesto por las mismas palabras ('la casa es bonita' vs ' la casa es bonita ').
- **Palabras faltantes**: si a un texto le faltan palabras con respecto al otro ('casa bonita' vs 'la casa es bonita').
- **Errores ortográficos**: si un texto presenta errores ortográficos como falta de tildes u otros errores ('información' vs 'informasion').
- **Errores tipográficos**: si un texto presenta caracteres sobrantes, faltantes o transpuestos con respecto al otro ('informción' vs 'informaciót')
- **Palabras truncadas**: si en un texto se usa abreviaturas o palabras truncadas ('B/quilla' vs 'Barranquilla')
- **Prefijos y sufijos**: si en un texto se usan prefijos o sufijos y en otro no ('Sr. Iván Amón' vs 'Iván Amón'). Nótese como este criterio no es el mismo que palabras faltantes, porque la palabra faltante puede presentarse en cualquier parte del texto, pero bajo este criterio, estaría al comienzo o al final de los textos.
- **Sinónimos**: si en un texto se usa una palabra que es sinónima de otra (´casa´ vs ´hogar´).

Los criterios deben aplicarse a las diferentes técnicas. La técnica de la distancia de edición estándar o distancia de Levehnstein [41], calcula la distancia existente entre dos textos como el número de operaciones de edición (inserciones, borrados y reemplazos) necesarias para transformar un texto en el otro. La tabla I resume el comportamiento de esta técnica para los criterios definidos.

De otra parte, deben definirse una serie de razonamientos que establezcan las condiciones para aplicar la técnica. Una guía metodológica completa y general, debe revisar múltiples aspectos. A continuación, se examinan algunos aspectos, a modo de ejemplo.

TABLA I. CRITERIOS DE EVALUACIÓN PARA LA TÉCNICA DISTANCIA DE EDICIÓN

Criterio	Eficacia
Palabras en orden	Alta
Mayúsculas/minúsculas	Baja
Espacios en blanco	Baja
Palabras faltantes	Baja
Errores ortográficos	Baja
Errores tipográficos	Baja
Palabras truncadas	Baja
Prefijos y sufijos	Baja
Sinónimos	Baja

Cuál es el tipo de datos?

Si el campo es de texto, aplican técnicas de detección de duplicados como la distancia de edición, pero si el campo es numérico los problemas a buscar pueden ser de datos atípicos (outliers) y se debe establecer la distribución de los datos antes de aplicar alguna fórmula.

¿Es el atributo analizado un nombre de una compañía?

Para este tipo de atributos, existen técnicas específicas como búsqueda en bases de datos del gobierno. Además, en estos casos, no será común que se varíe el orden de las palabras.

¿Es el atributo un nombre de persona?

En nombres, dependiendo de la forma de captura, puede ser común, que se varíe el orden de las palabras o que existan palabras truncadas.

¿Es una dirección?

Para atributos que almacenan direcciones, existen técnicas específicas relacionadas con georeferenciación.

¿Es un campo de contenido distinto a nombres o direcciones? (datos tipo texto).

En un campo de contenido diferente a estos, como el título de un proyecto de investigación, no será común que se varíe el orden de las palabras.

¿Se visualiza uso indistinto de mayúsculas y minúsculas?

En una muestra de datos, puede visualizarse si en los registros se usa indistintamente mayúsculas y minúsculas.

¿Se visualiza uso de prefijos y/o sufijos?

Un atributo que contenga nombres de clientes, seguramente no incluirá prefijos como Sr., Sra. o Dr. pero si se trata de uno que contenga nombres de investigadores, es posible que incluya prefijos o sufijos como MsC o PhD.

Estas y otras preguntas, organizadas en forma de árbol de decisión, conducirían a la selección de las técnicas adecuadas a un conjunto de datos particular. En este sentido, se dirigen nuestros esfuerzos.

V. CONCLUSIONES

Se han descrito algunos elementos importantes que deben ser tenidos en cuenta para la construcción de una guía metodológica que oriente la selección de las técnicas a ser aplicadas sobre un conjunto de datos particular, de acuerdo con las peculiaridades que éste presente.

Ya que los posibles problemas de los datos son muchos y la cantidad de técnicas existentes también es alta, la tarea de construir una guía metodológica completa que contemple los diferentes problemas y la mayor cantidad posible de técnicas existentes para detectarlos y corregirlos, es una tarea ardua pero que bien vale la pena para garantizar la correcta toma de decisiones.

REFERENCIAS

[1]Dasu, T., Vesonder, G. T., y Wright, J. R. 2003. Data quality through knowledge engineering. En: Proceedings of the Ninth International Conference on Knowledge Discovery and Data Mining ACM SIGKDD 2003 (Washington, D.C., Agosto 24 - 27, 2003). KDD '03. ACM, Nueva York, NY, 705-710. DOI= http://doi.acm.org/10.1145/956750.956844

[2]Gartner. Dirty Data is a Business Problem, Not an IT Problem. [En línea]. 2007. http://www.gartner.com/it/page.jsp?id=501733 [Consulta: Octubre 10 de 2008]

[3]Rahm, E., y Do, H. H. 2000. Data Cleaning: Problems and Current Approaches. IEEE Bulletin of the Technical Committee on Data Engineering, 24 (4).

[4]Kim, W., Choi, B.J., Hong, E.K., Kim, S.K., y Lee, D. 2003. A Taxonomy of Dirty Data. Data Mining and Knowledge Discovery, 7, 2003. 81-99.

[5]Müller, H., y Freytag, J.C. 2003. Problems, Methods, and Challenges in Comprehensive Data Cleansing. Technical Report HUB-IB-164, Humboldt University, Berlin.

[6]Oliveira, P., Rodrigues, F., Henriques, P., y Galhardas, H. 2005. A Taxonomy of Data Quality Problems. En: Second International Workshop on Data and Information Quality IQIS 2005 (Porto, Portugal, Junio 13-17, 2005).

[7]E-notes, Encyclopedia of Public Health. [En línea], http://www.enotes.com/public-health-encyclopedia/record-linkage [Consulta: Octubre 10 de 2008].

[8]Dunn, H. 1946. Record Linkage. American Journal of Public Health 36, 1412-1416.

[9]Newcombre, H.B., Kennedy, J.M., Axford, S., y James, A. 1959. Automatic linkage of vital records. Science. 130 (3381), 954-959, Octubre, 1959.

[10]Newcombe, H.B., Kennedy, J.M. 1962. Record Linkage: Making maximun use of the discriminating power of identifying information. Comm. ACM. 5(11), 563-566, Noviembre, 1962.

[11]Newcombe, H.B. 1967. Record Linking: The design of efficient systems for linking records into individual and family histories. Am. J. Human Genetics. 19 (3), 335-359, Mayo, 1967.

[12]Newcombe, H.B. 1988. Handbook of Record Linkage: Methods for health and statistical studies, administration, and business. Oxford: Oxford University Press.

[13]Newcombe, H.B., Fair, M.E., y Lalonde, P. 1992. The use of names for linking personal records. Journal of the American Statistical Association. 1193-1208.

[14]Winkler, W.E. 1988. Using the EM algorithm for Weight Computation in the Fellegi-Sunter model of Record Linkage. Proceedings of the Section on Survey Research Methods, American Statistical Association, 1988, 667-671.

[15]Winkler, W. E. 1989. Near Automatic Weight Computation in the Fellegi-Sunter Model of Record Linkage. En: Proceedings of the Fifth Census Bureau Annual Research Conference (Washington, D.C. Estados Unidos, Marzo 19-22, 1989), 145-155.

[16]Winkler, W. E. 1989. Methods for Adjusting for Lack of Independence in an Application of the Fellegi-Sunter Model of Record Linkage. Survey Methodology, 15, 101-117.

[17]Winkler, W. E. 1989, Frequency-based Matching in the Fellegi-Sunter Model of Record Linkage. En: Proceedings of the Section on Survey Research Methods, American Statistical Association,1989, 778-783.

[18]Winkler, W. E. 1990. Documentation of record-linkage software. unpublished report, Washington DC: Statistical Research Division, U.S. Bureau of the Census.

[19]Winkler, W. E. 1990. String Comparator Metrics y Enhanced Decision Rules in the Fellegi-Sunter Model of Record Linkage. En: Proceedings of the Section on Survey Research Methods,American Statistical Association, 1990, 354-359.

[20]Winkler, W. E. 1993. Business Name Parsing and Standardization Software. unpublished report, Washington, DC: Statistical Research Division, U.S. Bureau of the Census.

[21]Winkler, W. E. 1993. Improved Decision Rules in the Fellegi-Sunter Model of Record Linkage. En: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1993, 274-279.

[22]Winkler, W. E. 1994. Advanced Methods for Record Linkage. En: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1994, 467-472.

[23]Winkler, W. E. 1995. Matching and Record Linkage. in B. G. Cox et al. (ed.) Business Survey Methods, Nueva York: J. Wiley, 355-384.

[24]Winkler, W. E. y Scheuren, F. 1995. Linking Data to Create Information. En: Proceedings of Symposium 95, From Data to Information - Methods and Systems, Statistics Canada, 29-37.

[25]Winkler, W. E. y Scheuren, F. 1996. Recursive Analysis of Linked Data Files. En: Proceedings of the 1996 Census Bureau Annual Research Conference, 1996, 920-935.

[26]Winkler, W. E. 1997. Producing Public-Use Microdata That are Analytically Valid and Confidential. En: Proceedings of the Section on Survey Research Methods, American Statistical Association, 1997, 41-50.

[27]Winkler, W. E. 1998. Re-identification Methods for Evaluating the Confidentiality of Analytically Valid Microdata,_ Research in Official

Statistics, 1, 87-104.
[28]Winkler, W. E. 1999. Issues with Linking Files and Performing Analyses on the Merged Files. En: Proceedings of the Section on Social Statistics, American Statistical Association, 1999.
[29]Winkler, W. E. y Scheuren, F. 1991. How Computer Matching Error Effects Regression Analysis: Exploratory and Confirmatory Analysis, U.S. Bureau of the Census, Statistical Research Division Technical Report.
[30]Winkler, W. E. y Thibaudeau, Y. 1991. An Application of the Fellegi-Sunter Model of Record Linkage to the 1990 U.S. Census. U.S. Bureau of the Census, Statistical Research Division Technical report RR91/09.
[31]Elmagarmid, A., Ipeirotis, P., y Verykios, V. 2007. Duplicate Record Detection: A Survey. IEEE Transactions on knowledge and data engineering. 19 (1). Enero, 2007.
[32]Hancong L., Sirish S., y Wei, J. 2004. On-line outlier detection and data cleaning. Computers & chemical engineering, 2004, 28 (9), 1635-1647.
[33]Varun, Ch., Arindam, B., y Vipin K,. 2007. Outlier detection: A survey. Technical Report Department of Computer Science and Engineering. University of Minnesota. Agosto 15, 2007.
[34]Bakar, Z., Ahmad, M., y Deris, M. A. 2006. Comparative Study for Outlier Detection. En: IEEE Conference on Cybernetics and Intelligent Systems CIS 2006 (Bangkok, Tailandia, Junio 7-9, 2006).
[35]Matsumoto, S., Kamei, Y., Monden, A., y Matsumoto, K. 2007. Comparison of Outlier Detection Methods in Fault-proneness Models. En: Proceedings of the First international Symposium on Empirical Software Engineering and Measurement ESEM 2007 (Madrid, España, Septiembre 20 - 21, 2007). IEEE Computer Society, Washington, DC, 461-463. DOI= http://dx.doi.org/10.1109/ESEM.2007.34
[36]Angiulli, F., Basta, S., and Pizzuti, C. 2006. Distance-Based Detection and Prediction of Outliers. IEEE Trans. on Knowl. and Data Eng. 18, 2 (Feb. 2006), 145-160. DOI= http://dx.doi.org/10.1109/TKDE.2006.29
[37]Narita, K., y Kitagawa, H. 2008. Outlier Detection for Transaction Databases using Association Rules. En: Proceedings of the Ninth International Conference on Web-Age Information Management iiWAS2007, (Jakarta, Indonesia, Diciembre 3-5, 2007).
[38]Tierstein, Leslie. A Methodology for Data Cleansing and Conversion, White paper W R Systems, Ltd.
[39]Rosenthal, A., Wood, D., y Hughes, E. 2001. Methodology for Intelligence Database Data Quality. Julio, 2001.
[40]Rittman, Mark. 2006. Data Profiling and Automated Cleansing Using Oracle Warehouse Builder 10g Release 2, Septiembre, 2006.
[41]Ristad, E., y Yianilos, P. 1998. Learning string edit distance. 1998. IEEE Trans. Pattern Analysis and Machine Intelligence, 20(5), 522-532, Mayo, 1998.
[42]Waterman, M., Smith, y T., Beyer, W.A. 1976. Some biological sequence metrics. Advances in Math., 20(4), 367-387, 1976.
[43]Smith, T., y Waterman, M. 1981. Identification of common molecular subsequences. J. Molecular Biology, 147, 195-197.
[44]M.A. Jaro, 1976. Unimatch: A Record Linkage System: User's Manual, technical report, US Bureau of the Census, Washington, D.C.
[45]J.R. Ullmann, 1977. A Binary n-Gram Technique for Automatic Correction of Substitution, Deletion, Insertion, and Reversal Errors in Words. The Computer J., 20(2), 141-147.
[46] W.W. Cohen. 1998. Integration of Heterogeneous Databases without Common Domains Using Queries Based on Textual Similarity. En: Proceedings of the SIGMOD International Conference Management of Data SIGMOD'98 (Seattle, Washington, Junio 2-4, 1998), 201-212.
[47]R.C. Russell. 1918 Index, U.S. Patent 1,261,167, http://patft.uspto.gov/netahtml/srchnum.htm, Apr. 1918.
[48]R.C. Russell. 1922. Index, U.S. Patent 1,435,663, http://patft.uspto.gov/netahtml/srchnum.htm, Noviembre. 1922.
[49]R.L. Taft. 1970. Name Search Techniques. Technical Report Special Report No. 1, Nueva York State Identification and Intelligence System, Albany, N.Y., Febrero, 1970.
[50]L.E. Gill, 1997. OX-LINK: The Oxford Medical Record Linkage System. En: Proceedings of International Record Linkage Workshop and Exposition, (Arlington, Estados Unidos, Marzo 20-21, 1997), 15-33.
[51]L. Philips. 1990. Hanging on the Metaphone, Computer Language Magazine, 7(12), 39-44, Diciembre, 1990.
[52]L. Philips. 2000. The Double Metaphone Search Algorithm," C/C++ Users J., 18(5), Junio, 2000.
[53]S.A Jimenez-Marquez, C.Lacroix, y J. Thibault. 2002. Statistical data validation methods for large cheese plant database. J.Dairy Sci.,85(9), 2081-2097, Septiembre, 2002.
[54]M. M. Breuning,H. P. Kriegel, R. T. Ng, y J. Sander. 2000. Lof: Identifying density-based local outlier. En: Proceedings SIGMOD 2000 (Dallas, Texas, Mayo 14-19, 2000), 93-104.
[55]T. M. Khoshgoftaar, N. Seliya, y K. Gao. 2005. Detecting noisy instantes with the rule-based classification model. Intell. Data Anal., 9(4), 347-364, Julio 2005.
[56]E. Knorr and R. Ng, 1998. Algorithms for Mining Distance-Based Outliers. En: Large Datasets, Proc. Int'l Conf. Very Large Databases VLDB '98 (Nueva York, Estados Unidos, Agosto 24-27, 1998), 392-403.
[57]E. Knorr, R. Ng, y V. Tucakov. 2000. Distance-Based Outlier: Algorithms and Applications. VLDB J., 8(3-4), 237-253.
[58]J.A. Little y D. Rubin. 1989. Analysis of social science data with missing value. Sociological Methods Research.1989, 18, 292-326.
[59]Y.M. Babad, y J.A. Hoffer. 1984. Even no data has a value, Communications of the ACM, 27, 1984, 748-757
[60]Han, J. y M. Kamber, 2001. Data Mining: Concepts and Techniques, Morgan Kaufmann Publishers.
[61]CRISP-DM Consortium. CRISP-DM 1.0 Step-by-step data mining guide. Chicago, IL.: SPSS Inc., 2000, 13
[62]SAS. Enterprise Miner SEMMA [En línea]. Cary, NC: SAS Institute Inc., 2003. <http://www.sas.com/technologies/analytics/datamining/miner/semma.html> [Consulta: Octubre 10 de 2008]