Minería de Datos en Computación de Altas Prestaciones para Identificación en Grandes Bases de Datos de Huellas Dactilares

Daniel Peralta Cámara
Directores: Francisco Herrera Triguero
José Manuel Benítez Sánchez

Grupo de Investigación "Soft Computing and Intelligent Information Systems" (SCI^2S).
Departamento de Ciencias de la Computación e Inteligencia Artificial,
Universidad de Granada.
dperalta@decsai.ugr.es
Fecha de Inicio: Octubre de 2012

Palabras clave: Huellas Dactilares, Identificación, Grandes Bases de Datos, Sistemas Paralelos, Sistemas Distribuidos, Minería de Datos

1. Introducción

Uno de los problemas de mayor actualidad en la sociedad es la identificación de personas, presente en múltiples ámbitos. En este contexto la identificación mediante datos biométricos, y en particular utilizando huellas dactilares, cobra cada día mayor importancia [4]. Este problema puede abordarse desde dos perspectivas, cada una de los cuales constituye en sí misma un problema diferente [5]:

- **Verificación**: consiste en determinar si dos huellas dactilares corresponden a una misma persona.
- **Identificación**: consiste en, dada una base de datos que contenga información identificativa de diversas personas, determinar a cuál de ellas corresponde una determinada huella.

En general, para la verificación se extraen características de ambas huellas y se aplica un algoritmo de *matching* que determina su similitud [6].

En un problema de identificación, el proceso se repite para cada entrada de la base de datos hasta hallar una huella coincidente con la buscada. El porcentaje de la base de datos que se recorre (llamado *tasa de penetración*) es del 100 % en un sistema sin procesamiento específico. Por tanto estos métodos sólo son eficaces cuando el tamaño de la base de datos se limita a unos pocos individuos (desde decenas hasta miles dependiendo de los casos). Además, la precisión se deteriora cuando el número de huellas en la base de datos aumenta.

Mediante un sistema de computación de altas prestaciones (*High Performance Computing*, HPC) y adaptando el diseño y la implementación de los algoritmos, se puede distribuir el esfuerzo computacional entre distintos procesadores y máquinas para acelerar el proceso de identificación [7,1].

Por otra parte, de forma natural las huellas se dividen en cinco tipos morfológicos que se manifiestan en las siguientes proporciones: *Whorl* (27.9 %), *Right loop* (31.7 %), *Left loop* (33.8 %), *Arch* (3.7 %) y *Tented arch* (2.9 %). La tasa de penetración puede reducirse si una huella solamente se compara con las de su misma clase. Esto requiere un clasificador preciso, puesto que un error puede conllevar una identificación errónea o el recorrer toda la base de datos [2,3].

Finalmente, con el objetivo de maximizar la tasa de acierto en la identificación, surgen propuestas basadas en el uso simultáneo de huellas dactilares [8].

2. Hipótesis de Partida

En el momento de iniciar la tesis las propuestas en la literatura científica se centraban en bases de datos pequeñas, del orden de cientos o miles de huellas, por lo que carecían de la escalabilidad suficiente para grandes bases de datos. Aunque había sistemas basados en HPC, estaban orientados a objetivos diferentes como alta disponibilidad, servicio a dispositivos móviles o búsqueda remota. En cuanto a la clasificación, en general los algoritmos más precisos rechazan una parte de las huellas y no les asignan ninguna clase, aumentando la penetración. Finalmente, los sistemas basados en múltiples huellas mejoran la precisión de la verificación sin tener en cuenta el impacto en el tiempo.

Surge por tanto la necesidad de desarrollar técnicas y sistemas novedosos, capaces de identificar individuos en bases de datos muy grandes, con tasas de error mínimas y en un intervalo de tiempo limitado. Como punto de partida, se establecen las siguientes hipótesis:

- Los sistemas propuestos en la literatura no son capaces de satisfacer todas las necesidades de identificación de grandes instituciones.
- Con un sistema *software* adecuadamente diseñado sería posible mantener el mismo tiempo de búsqueda frente a un aumento del tamaño de la base de datos, aumentando el número de equipos en la arquitectura subyacente.
- La HPC debe emplearse de forma conjunta con una propuesta adecuada de nuevas metodologías de clasificación y extracción de características en la que no se rechace ninguna huella de entrada.
- Se pueden mejorar el tiempo y la precisión de la identificación con una búsqueda doble. Se recorre la base de datos con un algoritmo de *matching* rápido para extraer un conjunto de huellas similares a la de entrada, sobre el que se aplica un algoritmo de *matching* más complejo y preciso para determinar si la huella de entrada se encuentra o no entre ellas.
- La pérdida de precisión en grandes bases de datos puede reducirse utilizando dos huellas dactilares.
- Los nuevos paradigmas para Big Data son útiles para la identificación, especialmente cuando se involucran técnicas de extracción de características, clasificación, y varias huellas o algoritmos de *matching*. La aplicación de técnicas de reducción de dimensionalidad (como el *manifold learning*) puede ser un buen complemento.

Todas estas propuestas no son excluyentes, y es posible combinarlas en un único sistema final para así aprovechar todas sus ventajas. El problema de la identificación en bases de datos tan grandes no es trivial, y una solución escalable, precisa y adecuada requiere la combinación de varios enfoques diferentes.

3. Objetivos

En base a estas hipótesis se planteó como objetivo general el desarrollo de un sistema completo de identificación mediante huellas, que permita buscar en bases de datos muy grandes (del orden de millones o decenas de millones de huellas) en un espacio de tiempo razonable. Más concretamente, se consideraron los siguientes objetivos:

- **Sistema distribuido de búsqueda**: el principal cuello de botella en la identificación es la búsqueda secuencial. Por ello, se plantea el diseño y desarrollo de un sistema distribuido flexible basado en HPC, que posibilite la búsqueda en paralelo a través de distintas zonas de la bases de datos.
- **Identificación mediante doble huella**: para garantizar una tasa de acierto máxima, se propone la utilización de dos huellas dactilares para la identificación. De nuevo, un uso adecuado de la HPC se plantea como un factor clave para el desarrollo de un sistema eficiente.
- **Identificación en dos fases**: se propone el diseño y desarrollo de un sistema de identificación que utilice un algoritmo de *matching* muy rápido para extraer un conjunto de identidades candidatas, que son posteriormente procesadas por un algoritmo muy preciso.
- **Clasificación de huellas dactilares**: para reducir la tasa de penetración, se busca diseñar un clasificador con máximo acierto y mínimo rechazo, utilizando técnicas de Minería de Datos y combinación de clasificadores.

4. Metodología y Plan de Trabajo

Dada la necesidad de una metodología teórico-práctica, se requiere un método de trabajo basado en el método científico habitual y dé cabida a las necesidades de dicha metodología. En particular, el método seguido es el siguiente:

1. **Observación**: estudio del problema de la identificación biométrica mediante huellas dactilares y de sus características, así como de las posibilidades que ofrecen la HPC y las técnicas de minería de datos.
2. **Formulación de hipótesis**: diseño de nuevos métodos de identificación que usen algoritmos complementarios como la clasificación, capaces de manejar grandes bases de datos mediante HPC.
3. **Recogida de observaciones**: obtención de resultados tras la aplicación de los nuevos métodos a bases de datos de huellas.
4. **Contraste de hipótesis**: comparación de los resultados obtenidos con aquellos publicados en la literatura.

5. **Demostración o refutación de hipótesis**: aceptación o rechazo y modificación, si procede, de las técnicas desarrolladas tras las pruebas realizadas.
6. **Tesis o teoría científica**: extracción, redacción y aceptación de las conclusiones obtenidas durante el proceso.

Para alcanzar los objetivos siguiendo esta metodología, se definieron los siguientes pasos del plan de trabajo:

1. **Estudio pormenorizado del problema**: revisión de la literatura especializada en la identificación mediante huellas dactilares y las propuestas existentes para su resolución. Estudio de técnicas de paralelización y distribución orientadas a la resolución del problema, así como de técnicas del ámbito de la Minería de Datos para clasificación.
2. **Desarrollo de las propuestas**: implica además la validación de los resultados obtenidos comparándolos con los de otras técnicas consideradas referentes en la investigación actual.
   a) Diseño de un sistema distribuido (HPC) de identificación de huellas escalable que permita trabajar con grandes bases de datos.
   b) Diseño de algoritmos de *matching* e identificación con doble huella y en dos fases, para optimizar la precisión y el tiempo de la identificación.
   c) Diseño de un algoritmo de clasificación de huellas orientado a reducir la tasa de penetración en la base de datos.
   d) Integración de las anteriores ideas en un sistema completo de identificación eficiente, escalable y preciso al trabajar con grandes bases de datos.

5. Relevancia

No cabe duda de que el problema de identificación de personas es gran actualidad, y de vital importancia en diversos ámbitos. La propuesta de tesis pretende obtener un sistema directamente aplicable a casos reales, especialmente cuando impliquen bases de datos del orden de cientos de miles de personas en adelante.

A nivel científico, el trabajo desarrollado hasta el momento para esta tesis ha dado lugar a varias publicaciones en revistas internacionales de alto índice de impacto. Tres de ellas surgieron del profundo estudio realizado sobre el problema:

1. M. Galar, J. Derrac, D. Peralta, I. Triguero, D. Paternain, C. Lopez-Molina, S. García, J.M. Benítez, M. Pagola, E. Barrenechea, H. Bustince, F. Herrera. A Survey of Fingerprint Classification Part I: Taxonomies on Feature Extraction Methods and Learning Models. Knowledge-Based Systems 81 (2015) 76-97
2. M. Galar, J. Derrac, D. Peralta, I. Triguero, D. Paternain, C. Lopez-Molina, S. García, J.M. Benítez, M. Pagola, E. Barrenechea, H. Bustince, F. Herrera. A Survey of Fingerprint Classification Part II: Experimental Analysis and Ensemble Proposal. Knowledge-Based Systems 81 (2015) 98-116
3. D. Peralta, M. Galar, I. Triguero, D. Paternain, S. García, E. Barrenechea, J. M. Benítez, H. Bustince, F. Herrera. A Survey on Fingerprint Minutiae-based Local Matching for Verification and Identification: Taxonomy and Experimental Evaluation. Information Sciences 315 (2015) 67-87

Los avances en la vertiente práctica de la tesis han dado lugar al planteamiento de cuatro artículos (dos publicados, uno sometido y uno en desarrollo), demostrando la consecución de parte de los objetivos planteados:

1. D. Peralta, I. Triguero, R. Sanchez-Reillo, F. Herrera, J.M. Benítez. Fast Fingerprint Identification for Large Databases. Pattern Recognition 47:2 (2014) 588–602
   - Se propone un sistema de identificación paralelo a dos niveles, escalable y adaptable a cualquier algoritmo de *matching*, manteniendo su precisión.
2. D. Peralta, M. Galar, I. Triguero, O. Miguel-Hurtado, J.M. Benítez, F. Herrera. Minutiae Filtering to Improve Both Efficacy and Efficiency of Fingerprint Matching Algorithms. Engineering Applications of Artificial Intelligence, 32 (2014) 37-53
   - Se presenta un algoritmo de filtrado de minucias, basado en la segmentación de la imagen de la huella, para eliminar las minucias erróneamente detectadas en los bordes de la misma.
3. D. Peralta, I. Triguero, S. García, F. Herrera, J.M. Benítez. DPD-DFF: A Dual Phase Distributed Scheme with Double Fingerprint Fusion for Fast and Accurate Identification in Large Databases. (Submitted)
   - Se describe y evalúa un sistema con doble huella y doble *matching*.
4. D. Peralta, S. García, F. Herrera, J.M. Benítez. Fingerprint Identification in MapReduce and Spark. (Documento en preparación)
   - Se describen y comparan dos propuestas de identificación escalables sobre plataformas para Big Data.

El trabajo futuro se centra en la clasificación de huellas dactilares para la mejora de la identificación.

Referencias

1. Cappelli, R., Ferrara, M., Maltoni, D.: Large-scale fingerprint identification on GPU. Information Sciences 306, 1–20 (2015)
2. Galar, M., Derrac, J., Peralta, D., Triguero, I., Paternain, D., Lopez-Molina, C., García, S., Benítez, J., Pagola, M., Barrenechea, E., Bustince, H., Herrera, F.: A survey of fingerprint classification part I: Taxonomies on feature extraction methods and learning models. Knowledge-Based Systems 81, 76–97 (2015)
3. Galar, M., Derrac, J., Peralta, D., Triguero, I., Paternain, D., Lopez-Molina, C., García, S., Benítez, J., Pagola, M., Barrenechea, E., Bustince, H., Herrera, F.: A survey of fingerprint classification part II: Experimental analysis and ensemble proposal. Knowledge-Based Systems 81, 98–116 (2015)
4. Jain, A., Flynn, P., Ross, A.: Handbook of biometrics. Springer (2007)
5. Maltoni, D., Maio, D., Jain, A., Prabhakar, S.: Handbook of fingerprint recognition. Springer-Verlag New York Inc (2009)
6. Peralta, D., Galar, M., Triguero, I., Paternain, D., García, S., Barrenechea, E., Benitez, J., Bustince, H., Herera, F.: A survey on fingerprint minutiae-based local matching for verification and identification: Taxonomy and experimental evaluation. Information Sciences 315, 67–87 (2015)
7. Peralta, D., Triguero, I., Sanchez-Reillo, R., Herrera, F., Benitez, J.M.: Fast fingerprint identification for large databases. Pattern Recognition 47(2), 588–602 (2014)
8. Prabhakar, S., Jain, A.: Decision-level fusion in fingerprint verification. Pattern Recognition 35(4), 861–874 (2002)