

Sistemas de agrupamiento automático (clustering) en documentos mediante técnicas de softcomputing: Aplicaciones de algoritmos genéticos

Fecha: 30/07/2010

Ponente: José Raúl Fernández del Castillo Díez

Profesor Titular del Dpto. de Ciencias de la Computación.

Universidad de Alcalá

En pocos años la Web se ha convertido en una herramienta universal para todo tipo de actividades culturales, profesionales y comerciales. Los avances tecnológicos de los últimos años han provocado un aumento exponencial de la información disponible, lo que obliga al desarrollo de herramientas dedicadas a la gestión y Recuperación de Información (RI), entendida ésta como la selección de la información demandada por un usuario de entre toda la información disponible. Esta labor es realizada por los llamados sistemas de recuperación de información (SRI), una clase de sistemas de información que tratan con bases de datos compuestas por documentos y los índices de la información en ellos contenidos, que procesan las consultas de los usuarios y les entregan los documentos relevantes en un intervalo de tiempo apropiado.

La denominada “Hipótesis del Agrupamiento”, en la que “los documentos fuertemente asociados” tienden a ser relevantes para una misma consulta, nos permite agilizar el procesamiento documental al poder atender a las características de grupos de documentos en lugar de los documentos individuales. Con ello, al mismo tiempo que se logra una menor carga sobre el sistema, también mejora la eficacia y eficiencia el presentar al usuario los resultados en grupos de documentos similares con la información demandada.

Una de las líneas de investigación más activas es la Agrupación de Documentos (*Document Clustering*), entendida como la tarea de separar documentos en grupos afines según su similitud, en la que se buscan técnicas que permitan sustituir a aquellas basadas en el análisis de las relaciones completas (entre todos documentos) o los exitosos algoritmos basados en *k-means* y en redes neuronales.

Así, en los últimos años se ha experimentado sobre el paradigma de la Computación Evolutiva (Algoritmos Evolutivos, AE). Un AE funciona manteniendo una población de posibles soluciones del problema a resolver, llevando a cabo una serie de alteraciones sobre las mismas, y efectuando una selección para determinar cuáles permanecen en generaciones futuras y cuáles son eliminadas. Los AE ofrecen una metodología de búsqueda potente e independiente del dominio, aplicable a gran cantidad de tareas y que, en el caso de los Algoritmos Genéticos, se basan en operadores genéticos tales como el cruce y la mutación.

Los AE han sido aplicados con éxito para la solución de los problemas:

- Agrupamiento (clustering) de documentos y términos.
- Indización de documentos.
- Mejoras en la definición de consultas.
- Aprendizaje de funciones de similitud.

Durante la charla se introducirán nociones básicas de Algoritmos Genéticos para seguidamente tratar la aplicación de estos algoritmos al problema de la agrupación automática de documentos. Estudiaremos las técnicas de codificación, la definición de la población de individuos y de los individuos en sí mismos, así como los operadores genéticos de cruce y mutación. Para terminar se estudiará la técnica sobre un caso real de aplicación de AG al clustering automático.

Referencias Bibliográficas

Baeza-Yates R, Ribeiro-Neto B. “Modern Information Retrieval”. ACM Press Addison-Wesley, 1999. ISBN:0-201-39829-X .

Cordón O, Herrera-Viedma E. Zarco Carmen “A Review on the Application of Evolutionary Computation to Information Retrieval” Dept. of Computer Science and A.I University of Granada, Puleva Food S.A.

Kaufmann, L. y Rousseeuw, Peter J. 1990. "Finding Groups in data: An introduction to Cluster Analysis", John Wiley & Sons, Inc, NY.

Rijsbergen (Van) C.J. "Information Retrieval". Butterworth, London, 1979. Available at Computing Science. University of Glasgow.