

Índice general

1. Introducción	1
1.1. Objetivos	1
1.2. Estructura de la tesis	2
1.3. Estado del arte	2
1.3.1. Los sistemas de recuperación de información	3
1.3.2. Los paradigmas de la recuperación de información	4
1.3.3. El <i>clustering</i> en la recuperación de información	5
1.3.4. La familia del <i>K-Means</i>	7
1.3.5. La familia del DBSCAN	8
1.4. Justificación de objetivos	9
2. Los sistemas de recuperación de información	11
2.1. Aspectos generales	11
2.1.1. Qué vamos a entender por sistema de recuperación de información	11
2.1.2. ¿Para qué un sistema de recuperación de información?	12
2.1.3. Evolución de los sistemas de recuperación de información	13
2.1.4. Organización manual o automática de la información	14
2.1.5. Fases de un sistema de recuperación de información	15
2.1.6. Sistemas parecidos y relacionados	21
2.1.7. Problemática de los sistemas de recuperación de información	25
2.2. Preprocesamiento	32
2.2.1. Selección de los términos que identificarán la colección	32
2.2.2. Eliminación de palabras carentes de información: <i>Stopwords</i>	34
2.2.3. Trabajar únicamente con la raíz de las palabras: <i>Stemming</i>	35
2.2.4. Utilización de información sintáctica y/o semántica: Lematización	36
2.2.5. Reducciones heurísticas de términos poco o demasiado utilizados	36
2.2.6. Tesoros	37
2.3. Modelización	37
2.3.1. Clasificación de los modelos	37
2.3.2. Modelo vectorial básico: Matriz de pesos	40
2.3.3. Modelo de indexación semántica latente (LSI): Mediante la SVD	46
2.3.4. Modelo de clustering	50
2.4. Tratamiento de la consulta	52

2.4.1.	Tipos de consultas	52
2.4.2.	Retroalimentación de la consulta	53
2.5.	Métodos de evaluación	55
2.5.1.	Evaluación específica de clusters	56
2.5.2.	Medidas de evaluación estándar	57
2.5.3.	Otras medidas de evaluación menos comunes	60
2.5.4.	Colecciones de prueba para evaluación	62
2.5.5.	Colecciones de prueba utilizadas	62
2.6.	Conclusiones	64
3.	Algoritmos de <i>clustering</i>	71
3.1.	Tipos de métodos de <i>clustering</i>	71
3.1.1.	<i>Clustering</i> particional	73
3.1.2.	<i>Clustering</i> jerárquico	74
3.1.3.	<i>Clustering</i> basado en densidades	75
3.2.	Métodos básicos	77
3.2.1.	<i>K-Means</i>	77
3.2.2.	DBSCAN	79
3.3.	Variante del <i>K-Means</i> : α - <i>Bisecting Spherical K-Means</i>	82
3.3.1.	α - <i>Bisection</i>	82
3.3.2.	α - <i>Bisecting Spherical K-Means</i>	84
3.4.	Variante del DBSCAN: VDBSCAN	85
3.4.1.	El algoritmo VDBSCAN	86
3.4.2.	Estudio de prestaciones de recuperación	87
3.4.3.	Estudio temporal	89
3.5.	Eliminación de parámetros del VDBSCAN	90
3.5.1.	Fijación del mínimo número de elementos por <i>cluster</i>	93
3.5.2.	Heurística para seleccionar una buena proximidad	95
3.6.	Conclusiones	99
4.	Clustering en sistemas distribuidos	105
4.1.	Introducción	105
4.2.	Posibles distribuciones	107
4.3.	α - <i>Bisecting Spherical K-Means</i> distribuido	109
4.3.1.	α - <i>Bisecting K-Means</i>	109
4.3.2.	α - <i>Bisecting Spherical K-Means</i>	113
4.4.	VDBSCAN distribuido	114
4.4.1.	Versiones paralelas del DBSCAN	114
4.4.2.	Versión distribuida del VDBSCAN	116
4.5.	Evaluación distribuida de consultas	118
4.6.	Conclusiones	119

5. Estudios experimentales	123
5.1. Colecciones de prueba y entornos utilizados	123
5.1.1. Colección: <i>Times Magazine</i> 1963	123
5.1.2. Colección: <i>Cystic Fibrosis Database</i>	123
5.1.3. Colección: TREC-DOE	124
5.1.4. <i>Cluster</i> de PCs: KEFREN	124
5.1.5. <i>Cluster</i> de PCs: ODIN	124
5.2. Comparación de calidad de recuperación	124
5.2.1. Matriz de Pesos vs. <i>Spherical K-Means</i>	125
5.2.2. Matriz de Pesos vs. α - <i>Bisecting Spherical K-Means</i>	127
5.2.3. <i>Spherical K-Means</i> vs. α - <i>Bisecting Spherical K-Means</i>	130
5.2.4. Matriz de Pesos vs. VDBSCAN	131
5.2.5. α - <i>Bisecting Spherical K-Means</i> vs. VDBSCAN	133
5.3. Comparación de prestaciones computacionales	134
5.3.1. Modelización: VDBSCAN vs. α - <i>Bisecting Spherical K-Means</i> vs. <i>Spherical K-Means</i>	136
5.3.2. Evaluación: Matriz de Pesos vs. <i>Spherical K-Means</i> vs. α - <i>Bisecting</i> <i>Spherical K-Means</i>	140
5.3.3. Evaluación: Matriz de Pesos vs. VDBSCAN	145
5.3.4. Evaluación: VDBSCAN vs. α - <i>Bisecting Spherical K-Means</i>	149
5.4. Conclusiones	153
5.4.1. Prestaciones en recuperación de información	154
5.4.2. Prestaciones computacionales	154
5.4.3. Mejor método según sus prestaciones	156
6. Conclusiones finales y trabajos futuros	159
6.1. Evolución de la tesis y producción científica	159
6.1.1. Evolución de la tesis	159
6.1.2. Publicaciones de la tesis	164
6.2. Conclusiones	168
6.3. Trabajos Futuros	170
Bibliografía	173