Capítulo Quinto

Análisis de Resultados

5. Análisis de resultados

Retomando parte del proceso experimental descrito en la sección 4.2. de la presente tesis, a continuación se presenta el análisis y resultados obtenidos sobre el desempeño y eficiencia del algoritmo ActiveRank como método para la clasificación automática de información.

5.1. Acerca de la interpretación visual de la matriz de rankings ActiveRank

Debido al enorme tamaño de la matriz de rankings generada por ActiveRank, su manipulación e interpretación en forma tabular puede ser extremadamente compleja en aplicaciones de una escala relativamente pequeña, una alternativa útil en el proceso de revisión y evaluación es representar la matriz en una gráfica tridimensional, donde los ejes *x* y *y* representan la combinación de los elementos *i* y *j*, y el eje *z* representa el valor de ranking *R(i,j)*. Cuando se trata de analizar el comportamiento del ranking de un elemento contra todos los demás se puede graficar bidimensionalmente la fila de interés de la matriz, y de esta forma analizar gráficamente los resultados obtenidos, método ampliamente utilizado en la presente tesis.

5.2. Proceso de clasificación manual de la muestra

La siguiente tabla despliega la evaluación manual del tipo de página para cada uno de los vectores pertenecientes a la muestra del 2% como se especificó en la sección 4.2. de esta tesis. Se muestra el identificador del vector, seguido por el ranking de dicho elemento contra el vector pornográfico de referencia, y por último VERDADERO si la página es pornográfica, y FALSO si el documento no es pornográfico. La clasificación se realizó abriendo el URL de cada uno de los documentos especificados en un navegador web y observando su contenido.

Tabla 5.2.1. – Clasificación manual de la muestra generada aleatoriamente mostrando el identificador del vector, su valor de ranking con respecto al vector pornográfico de referencia y la condición booleana para la premisa de pertenencia al conjunto de interés.

id_vector	ranking	real type
108	0.716147	VERDADERO
119	0.702886	VERDADERO
151	0.701142	VERDADERO
210	0.715083	VERDADERO
246	0.715083	VERDADERO
257	0.716821	VERDADERO
409	1.3208	VERDADERO
407	1.34164	VERDADERO
427	1.33924	VERDADERO
502	1.34816	VERDADERO
499	1.32817	VERDADERO
556	0.629866	VERDADERO
538	0.637565	VERDADERO
738	0.627553	VERDADERO
724	0.626547	VERDADERO
811	0.6284	VERDADERO
775	0.62727	VERDADERO
849	1.49506	VERDADERO
999	1.88974	VERDADERO
987	1.85913	VERDADERO
1091	1.4369	VERDADERO
1126	1.4369	VERDADERO
1125	1.79746	VERDADERO
1358	1.68373	VERDADERO
1392	1.4142	VERDADERO
1386	1.4142	VERDADERO
1333	1.87804	FALSO
1533	1.48021	VERDADERO
1641	1.47048	VERDADERO
1612	1.3867	VERDADERO
1623	1.48021	VERDADERO
1565	1.50475	VERDADERO
1869	1.50129	VERDADERO
1833	1.4481	VERDADERO

id_vector	ranking	real type
3369	1.90752	FALSO
3826	1.82897	FALSO
3825	2	FALSO
3838	1.88287	FALSO
3811	1.92608	FALSO
3683	1.93589	FALSO
4081	1.87545	FALSO
4046	1.98123	FALSO
3973	1.86834	FALSO
3996	2	FALSO
3934	1.86464	FALSO
4117	1.88798	FALSO
4126	1.90143	FALSO
4213	1.86834	FALSO
4165	1.85013	FALSO
4264	1.88273	FALSO
4261	1.87517	FALSO
4273	1.8582	FALSO
4915	1.89964	FALSO
4896	1.8607	FALSO
5027	1.92712	FALSO
5113	1.89659	FALSO
4726	1.93094	FALSO
4774	1.93589	FALSO
4802	1.88287	FALSO
4847	1.85729	FALSO
5595	1.92103	FALSO
5542	1.86182	FALSO
5220	1.93082	FALSO
5235	1.86099	FALSO
5232	1.91361	FALSO
5344	1.90752	FALSO
5373	1.94568	FALSO
5270	1.93589	FALSO

id_vector	ranking	real type
8150	1.87968	FALSO
8085	1.91733	FALSO
8116	1.8298	FALSO
7977	1.9046	FALSO
7984	1.87789	FALSO
7985	1.88842	FALSO
7894	1.91733	FALSO
7803	2	FALSO
8807	1.93589	FALSO
8791	2	FALSO
8765	1.84861	FALSO
8720	1.93589	FALSO
8895	1.91733	FALSO
9065	1.98961	FALSO
9010	1.94178	FALSO
9206	1.98123	FALSO
9180	1.94336	FALSO
9148	1.98931	FALSO
8274	1.91064	FALSO
8306	1.88734	FALSO
8208	1.9323	FALSO
8372	1.90354	FALSO
8551	1.93597	FALSO
8468	1.89977	FALSO
8496	1.89828	FALSO
8697	1.9046	FALSO
8687	2	FALSO
8585	1.9046	FALSO
9806	2	FALSO
9863	1.90342	FALSO
9868	1.87689	FALSO
9943	1.93302	FALSO
10003	1.91733	FALSO
10144	1.94517	FALSO

1793	1.48021	VERDADERO
2021	0.630998	VERDADERO
1970	0.637044	VERDADERO
1966	0.638366	VERDADERO
2208	0.638558	VERDADERO
2300	1.15942	VERDADERO
2296	1.15183	VERDADERO
2111	0.626995	VERDADERO
2085	0.630548	VERDADERO
2174	0.636286	VERDADERO
2477	1.15031	VERDADERO
2505	1.55085	VERDADERO
2365	1.16183	VERDADERO
2692	1.15393	VERDADERO
2669	1.16103	VERDADERO
2653	0.950767	VERDADERO
3033	1.89384	FALSO
2875	1.1562	VERDADERO
2929	1.1628	VERDADERO
2939	1.16344	VERDADERO
2899	1.46988	VERDADERO
3319	1.88287	FALSO
3220	1.93303	FALSO
3094	1.82897	FALSO
3126	1.92103	FALSO
3107	1.88287	FALSO
3582	1.97198	FALSO
3504	1.83989	FALSO
3410	1.90732	FALSO
3337	1.94917	FALSO
3347	1.88157	FALSO

5956	1.90864	FALSO
5984	1.82101	FALSO
5932	1.87654	FALSO
6120	1.92021	FALSO
6020	1.91733	FALSO
6026	1.91239	FALSO
5716	1.89266	FALSO
5749	1.91629	FALSO
5859	1.93082	FALSO
5762	1.94568	FALSO
6564	1.91066	FALSO
6634	1.91066	FALSO
6647	1.89322	FALSO
6603	1.91066	FALSO
6445	1.94262	FALSO
6526	1.91066	FALSO
6332	1.91066	FALSO
6314	1.91066	FALSO
6346	1.89547	FALSO
6165	1.93313	FALSO
7059	1.91024	FALSO
7125	1.92712	FALSO
7147	1.92712	FALSO
6914	1.92712	FALSO
6930	1.94178	FALSO
6746	1.91066	FALSO
7617	1.9196	FALSO
7391	1.88672	FALSO
7273	1.91881	FALSO
7249	1.92712	FALSO
7212	1.92712	FALSO

10231	2	FALSO
10197	1.94056	FALSO
10181	2	FALSO
9360	2	FALSO
9435	2	FALSO
9441	1.88524	FALSO
9462	2	FALSO
9459	1.91733	FALSO
9470	1.88287	FALSO
9488	1.93209	FALSO
9557	1.89669	FALSO
9625	1.86434	FALSO
9628	2	FALSO
9609	1.9046	FALSO
9613	1.84604	FALSO
9652	1.90247	FALSO
9670	1.89019	FALSO
9722	2	FALSO
9707	1.89948	FALSO
10445	1.83683	FALSO
10481	2	FALSO
10467	1.96184	FALSO
10426	1.97059	FALSO
10311	1.92922	FALSO
10359	1.89234	FALSO
10263	1.89773	FALSO
10297	1.94508	FALSO
10700	1.86661	FALSO
10499	1.89885	FALSO
10531	1.92296	FALSO

5.3. Medición de la eficiencia del proceso de clasificación

El proceso de clasificación automática, como se ha mencionado a lo largo de la presente tesis, consiste en discriminar a través del valor de ranking si un documento o información pertenece o no a un conjunto determinado; el método para lograr determinar el umbral óptimo de clasificación consiste en un

proceso de maximización de la eficiencia de clasificación en una muestra representativa del universo. Siguiendo el desarrollo experimental planteado en la sección 4.2. de la presente tesis, el valor inicial para el umbral de clasificación corresponde al valor medio entre $\bar{r}_{porno}$ y $\bar{r}_{católico}$ expresados en la tabla 4.3.1., y que corresponde a $r_{umbral\ inicial} = 1.53224907$; la tabla 5.3.1. describe para cada uno de los elementos de la muestra su identificador, el valor de ranking de dicho elemento contra el vector de referencia pornográfico, su pertenencia o no pertenencia al conjunto determinado manualmente como valor de control, su pertenencia o no pertenencia según el algoritmo de clasificación automática con respecto al valor $r_{umbral\ inicial}$, y la matriz binaria de calificación de *Verdadero-Verdadero, Falso-Verdadero, Verdadero-Falso* y *Falso-Falso* para cada una de las hipótesis de clasificación automática, la cual se construye a partir del siguiente algoritmo presentado en pseudocódigo:

SI *el documento es autoclasificado como pornográfico* Y *el documento es pornográfico* ENTONCES: *Verdadero-Verdadero*

SI *el documento es autoclasificado como pornográfico* Y *el documento no es pornográfico* ENTONCES: *Falso-Verdadero*

SI *el documento es autoclasificado como no pornográfico* Y *el documento no es pornográfico* ENTONCES: *Verdadero-Falso*

SI *el documento es autoclasificado como no pornográfico* Y *el documento es pornográfico* ENTONCES: *Falso-Falso*

NOTA: Si el algoritmo anterior se desea programar con estructuras *IF* anidadas, la condición inicial deberá ser respecto a la evaluación manual, y en término secundario la comparación de autoclasificación, de lo contrario se excluirá una rama completa del árbol (observar con condiciones de frontera).

SI *el documento es pornográfico* ENTONCES:
    SI *el documento es autoclasificado como pornográfico* ENTONCES: *Verdadero-Verdadero*
    SI NO: *Falso-Verdadero*
SI NO:
    SI *el documento es autoclasificado como pornográfico* ENTONCES: *Falso-Falso*
    SI NO: *Verdadero-Falso*

Tabla 5.3.1. – Tabla que despliega los resultados del análisis y proceso de autoclasificación de la etapa 1 del proceso experimental.

		$r_{umbral\ inicial}$	1.53224907	CLASIFIED AS PORN		CLASIFIED AS NOT PORN	
id_vector	**ranking**	**real type**	**auto clasification**	**TrueTrue**	**FalseTrue**	**TrueFalse**	**FalseFalse**
108	0.716147	VERDADERO	VERDADERO	1	0	0	0
119	0.702886	VERDADERO	VERDADERO	1	0	0	0
151	0.701142	VERDADERO	VERDADERO	1	0	0	0
210	0.715083	VERDADERO	VERDADERO	1	0	0	0
246	0.715083	VERDADERO	VERDADERO	1	0	0	0
257	0.716821	VERDADERO	VERDADERO	1	0	0	0
409	1.3208	VERDADERO	VERDADERO	1	0	0	0
407	1.34164	VERDADERO	VERDADERO	1	0	0	0
427	1.33924	VERDADERO	VERDADERO	1	0	0	0
502	1.34816	VERDADERO	VERDADERO	1	0	0	0
499	1.32817	VERDADERO	VERDADERO	1	0	0	0
556	0.629866	VERDADERO	VERDADERO	1	0	0	0
538	0.637565	VERDADERO	VERDADERO	1	0	0	0
738	0.627553	VERDADERO	VERDADERO	1	0	0	0
724	0.626547	VERDADERO	VERDADERO	1	0	0	0
811	0.6284	VERDADERO	VERDADERO	1	0	0	0
775	0.62727	VERDADERO	VERDADERO	1	0	0	0
849	1.49506	VERDADERO	VERDADERO	1	0	0	0
999	1.88974	VERDADERO	FALSO	0	0	0	1
987	1.85913	VERDADERO	FALSO	0	0	0	1
1091	1.4369	VERDADERO	VERDADERO	1	0	0	0
1126	1.4369	VERDADERO	VERDADERO	1	0	0	0
1125	1.79746	VERDADERO	FALSO	0	0	0	1
1358	1.68373	VERDADERO	FALSO	0	0	0	1
1392	1.4142	VERDADERO	VERDADERO	1	0	0	0
1386	1.4142	VERDADERO	VERDADERO	1	0	0	0
1333	1.87804	FALSO	FALSO	0	0	1	0
1533	1.48021	VERDADERO	VERDADERO	1	0	0	0
1641	1.47048	VERDADERO	VERDADERO	1	0	0	0
1612	1.3867	VERDADERO	VERDADERO	1	0	0	0
1623	1.48021	VERDADERO	VERDADERO	1	0	0	0
1565	1.50475	VERDADERO	VERDADERO	1	0	0	0
1869	1.50129	VERDADERO	VERDADERO	1	0	0	0

1833	1.4481	VERDADERO	VERDADERO	1	0	0	0
1793	1.48021	VERDADERO	VERDADERO	1	0	0	0
2021	0.630998	VERDADERO	VERDADERO	1	0	0	0
1970	0.637044	VERDADERO	VERDADERO	1	0	0	0
1966	0.638366	VERDADERO	VERDADERO	1	0	0	0
2208	0.638558	VERDADERO	VERDADERO	1	0	0	0
2300	1.15942	VERDADERO	VERDADERO	1	0	0	0
2296	1.15183	VERDADERO	VERDADERO	1	0	0	0
2111	0.626995	VERDADERO	VERDADERO	1	0	0	0
2085	0.630548	VERDADERO	VERDADERO	1	0	0	0
2174	0.636286	VERDADERO	VERDADERO	1	0	0	0
2477	1.15031	VERDADERO	VERDADERO	1	0	0	0
2505	1.55085	VERDADERO	FALSO	0	0	0	1
2365	1.16183	VERDADERO	VERDADERO	1	0	0	0
2692	1.15393	VERDADERO	VERDADERO	1	0	0	0
2669	1.16103	VERDADERO	VERDADERO	1	0	0	0
2653	0.950767	VERDADERO	VERDADERO	1	0	0	0
3033	1.89384	FALSO	FALSO	0	0	1	0
2875	1.1562	VERDADERO	VERDADERO	1	0	0	0
2929	1.1628	VERDADERO	VERDADERO	1	0	0	0
2939	1.16344	VERDADERO	VERDADERO	1	0	0	0
2899	1.46988	VERDADERO	VERDADERO	1	0	0	0
3319	1.88287	FALSO	FALSO	0	0	1	0
3220	1.93303	FALSO	FALSO	0	0	1	0
3094	1.82897	FALSO	FALSO	0	0	1	0
3126	1.92103	FALSO	FALSO	0	0	1	0
3107	1.88287	FALSO	FALSO	0	0	1	0
3582	1.97198	FALSO	FALSO	0	0	1	0
3504	1.83989	FALSO	FALSO	0	0	1	0
3410	1.90732	FALSO	FALSO	0	0	1	0
3337	1.94917	FALSO	FALSO	0	0	1	0
3347	1.88157	FALSO	FALSO	0	0	1	0
3369	1.90752	FALSO	FALSO	0	0	1	0
3826	1.82897	FALSO	FALSO	0	0	1	0
3825	2	FALSO	FALSO	0	0	1	0
3838	1.88287	FALSO	FALSO	0	0	1	0
3811	1.92608	FALSO	FALSO	0	0	1	0
3683	1.93589	FALSO	FALSO	0	0	1	0

4081	1.87545	FALSO	FALSO	0	0	1	0
4046	1.98123	FALSO	FALSO	0	0	1	0
3973	1.86834	FALSO	FALSO	0	0	1	0
3996	2	FALSO	FALSO	0	0	1	0
3934	1.86464	FALSO	FALSO	0	0	1	0
4117	1.88798	FALSO	FALSO	0	0	1	0
4126	1.90143	FALSO	FALSO	0	0	1	0
4213	1.86834	FALSO	FALSO	0	0	1	0
4165	1.85013	FALSO	FALSO	0	0	1	0
4264	1.88273	FALSO	FALSO	0	0	1	0
4261	1.87517	FALSO	FALSO	0	0	1	0
4273	1.8582	FALSO	FALSO	0	0	1	0
4915	1.89964	FALSO	FALSO	0	0	1	0
4896	1.8607	FALSO	FALSO	0	0	1	0
5027	1.92712	FALSO	FALSO	0	0	1	0
5113	1.89659	FALSO	FALSO	0	0	1	0
4726	1.93094	FALSO	FALSO	0	0	1	0
4774	1.93589	FALSO	FALSO	0	0	1	0
4802	1.88287	FALSO	FALSO	0	0	1	0
4847	1.85729	FALSO	FALSO	0	0	1	0
5595	1.92103	FALSO	FALSO	0	0	1	0
5542	1.86182	FALSO	FALSO	0	0	1	0
5220	1.93082	FALSO	FALSO	0	0	1	0
5235	1.86099	FALSO	FALSO	0	0	1	0
5232	1.91361	FALSO	FALSO	0	0	1	0
5344	1.90752	FALSO	FALSO	0	0	1	0
5373	1.94568	FALSO	FALSO	0	0	1	0
5270	1.93589	FALSO	FALSO	0	0	1	0
5956	1.90864	FALSO	FALSO	0	0	1	0
5984	1.82101	FALSO	FALSO	0	0	1	0
5932	1.87654	FALSO	FALSO	0	0	1	0
6120	1.92021	FALSO	FALSO	0	0	1	0
6020	1.91733	FALSO	FALSO	0	0	1	0
6026	1.91239	FALSO	FALSO	0	0	1	0
5716	1.89266	FALSO	FALSO	0	0	1	0
5749	1.91629	FALSO	FALSO	0	0	1	0
5859	1.93082	FALSO	FALSO	0	0	1	0
5762	1.94568	FALSO	FALSO	0	0	1	0

6564	1.91066	FALSO	FALSO	0	0	1	0
6634	1.91066	FALSO	FALSO	0	0	1	0
6647	1.89322	FALSO	FALSO	0	0	1	0
6603	1.91066	FALSO	FALSO	0	0	1	0
6445	1.94262	FALSO	FALSO	0	0	1	0
6526	1.91066	FALSO	FALSO	0	0	1	0
6332	1.91066	FALSO	FALSO	0	0	1	0
6314	1.91066	FALSO	FALSO	0	0	1	0
6346	1.89547	FALSO	FALSO	0	0	1	0
6165	1.93313	FALSO	FALSO	0	0	1	0
7059	1.91024	FALSO	FALSO	0	0	1	0
7125	1.92712	FALSO	FALSO	0	0	1	0
7147	1.92712	FALSO	FALSO	0	0	1	0
6914	1.92712	FALSO	FALSO	0	0	1	0
6930	1.94178	FALSO	FALSO	0	0	1	0
6746	1.91066	FALSO	FALSO	0	0	1	0
7617	1.9196	FALSO	FALSO	0	0	1	0
7391	1.88672	FALSO	FALSO	0	0	1	0
7273	1.91881	FALSO	FALSO	0	0	1	0
7249	1.92712	FALSO	FALSO	0	0	1	0
7212	1.92712	FALSO	FALSO	0	0	1	0
8150	1.87968	FALSO	FALSO	0	0	1	0
8085	1.91733	FALSO	FALSO	0	0	1	0
8116	1.8298	FALSO	FALSO	0	0	1	0
7977	1.9046	FALSO	FALSO	0	0	1	0
7984	1.87789	FALSO	FALSO	0	0	1	0
7985	1.88842	FALSO	FALSO	0	0	1	0
7894	1.91733	FALSO	FALSO	0	0	1	0
7803	2	FALSO	FALSO	0	0	1	0
8807	1.93589	FALSO	FALSO	0	0	1	0
8791	2	FALSO	FALSO	0	0	1	0
8765	1.84861	FALSO	FALSO	0	0	1	0
8720	1.93589	FALSO	FALSO	0	0	1	0
8895	1.91733	FALSO	FALSO	0	0	1	0
9065	1.98961	FALSO	FALSO	0	0	1	0
9010	1.94178	FALSO	FALSO	0	0	1	0
9206	1.98123	FALSO	FALSO	0	0	1	0
9180	1.94336	FALSO	FALSO	0	0	1	0

9148	1.98931	FALSO	FALSO	0	0	1	0
8274	1.91064	FALSO	FALSO	0	0	1	0
8306	1.88734	FALSO	FALSO	0	0	1	0
8208	1.9323	FALSO	FALSO	0	0	1	0
8372	1.90354	FALSO	FALSO	0	0	1	0
8551	1.93597	FALSO	FALSO	0	0	1	0
8468	1.89977	FALSO	FALSO	0	0	1	0
8496	1.89828	FALSO	FALSO	0	0	1	0
8697	1.9046	FALSO	FALSO	0	0	1	0
8687	2	FALSO	FALSO	0	0	1	0
8585	1.9046	FALSO	FALSO	0	0	1	0
9806	2	FALSO	FALSO	0	0	1	0
9863	1.90342	FALSO	FALSO	0	0	1	0
9868	1.87689	FALSO	FALSO	0	0	1	0
9943	1.93302	FALSO	FALSO	0	0	1	0
10003	1.91733	FALSO	FALSO	0	0	1	0
10144	1.94517	FALSO	FALSO	0	0	1	0
10231	2	FALSO	FALSO	0	0	1	0
10197	1.94056	FALSO	FALSO	0	0	1	0
10181	2	FALSO	FALSO	0	0	1	0
9360	2	FALSO	FALSO	0	0	1	0
9435	2	FALSO	FALSO	0	0	1	0
9441	1.88524	FALSO	FALSO	0	0	1	0
9462	2	FALSO	FALSO	0	0	1	0
9459	1.91733	FALSO	FALSO	0	0	1	0
9470	1.88287	FALSO	FALSO	0	0	1	0
9488	1.93209	FALSO	FALSO	0	0	1	0
9557	1.89669	FALSO	FALSO	0	0	1	0
9625	1.86434	FALSO	FALSO	0	0	1	0
9628	2	FALSO	FALSO	0	0	1	0
9609	1.9046	FALSO	FALSO	0	0	1	0
9613	1.84604	FALSO	FALSO	0	0	1	0
9652	1.90247	FALSO	FALSO	0	0	1	0
9670	1.89019	FALSO	FALSO	0	0	1	0
9722	2	FALSO	FALSO	0	0	1	0
9707	1.89948	FALSO	FALSO	0	0	1	0
10445	1.83683	FALSO	FALSO	0	0	1	0
10481	2	FALSO	FALSO	0	0	1	0

10467	1.96184	FALSO	FALSO	0	0	1	0
10426	1.97059	FALSO	FALSO	0	0	1	0
10311	1.92922	FALSO	FALSO	0	0	1	0
10359	1.89234	FALSO	FALSO	0	0	1	0
10263	1.89773	FALSO	FALSO	0	0	1	0
10297	1.94508	FALSO	FALSO	0	0	1	0
10700	1.86661	FALSO	FALSO	0	0	1	0
10499	1.89885	FALSO	FALSO	0	0	1	0
10531	1.92296	FALSO	FALSO	0	0	1	0

Suma Columnas	**48**	**0**	**141**	**5**

Una vez obtenida la matriz de aciertos en el proceso de autoclasificación se procede a calcular los coeficientes de eficiencia del proceso, a partir del número total de repeticiones de cada caso para un valor de umbral determinado; dichos datos se presentan en la última fila de la tabla 5.3.1..

Los índices de eficiencia se calculan como sigue:

$$E_{\in} = \frac{\sum VerdaderoVerdadero}{\sum VerdaderoVerdadero + \sum FalsoVerdadero}$$

$$E_{\notin} = \frac{\sum VerdaderoFalso}{\sum VerdaderoFalso + \sum FalsoFalso}$$

Donde $E_{\in}$ representa la eficiencia del método al clasificar elementos como pertenecientes al conjunto deseado, y $E_{\notin}$ la eficiencia del método al clasificar elementos como no pertenecientes al conjunto deseado, ambos utilizando un valor de umbral determinado.

Con $r_{umbral\ inicial} = 1.53224907$, los dos valores de eficiencia obtenidos fueron:

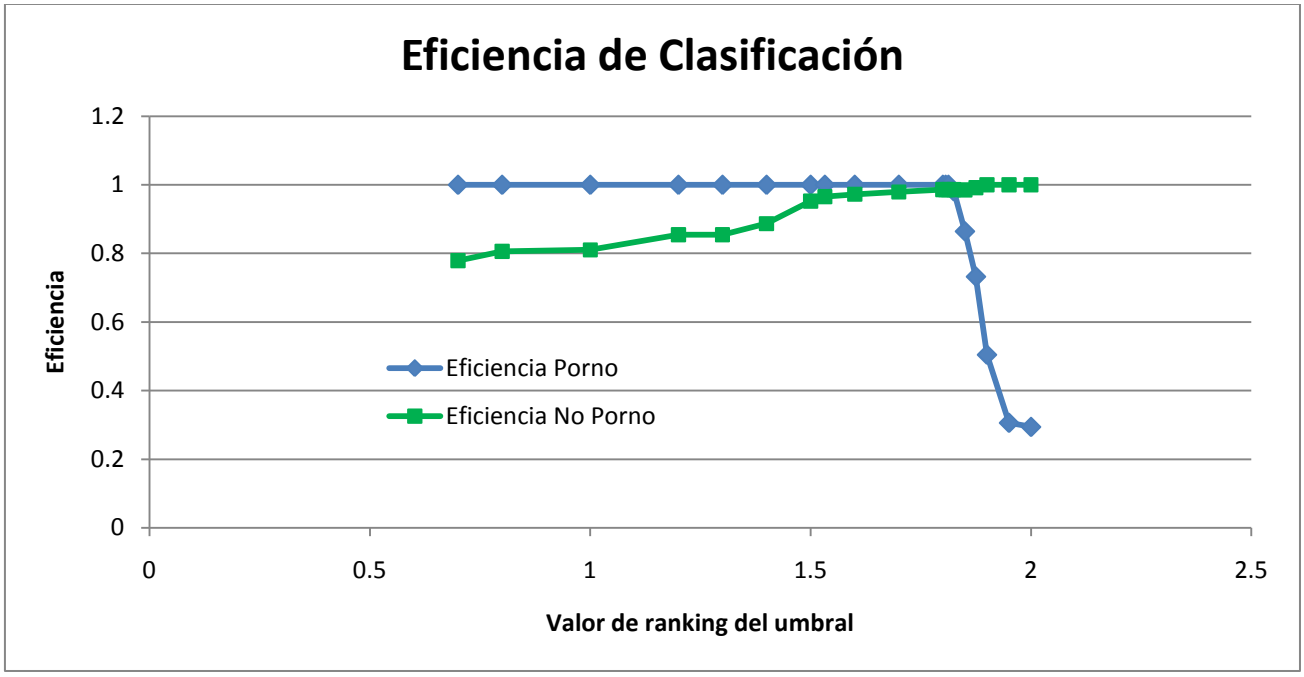
$$E_{\in} = 100\% \quad y \quad E_{\notin} = 96.57\%$$

Para realizar un análisis más detallado del comportamiento de la eficiencia en el proceso de autoclasificación, y que además permite plantear las bases de un sistema semiautomático para la obtención del valor de ranking umbral óptimo de mejor manera, se repitió el procedimiento planteado anteriormente para diferentes valores de umbral, inicialmente divididos en intervalos de 0.1, y aproximando por mitades el punto óptimo (se encuentra el intervalo que presenta máxima eficiencia, se divide en dos mitades iguales, se calcula de

nuevo la eficiencia en el subconjunto que presenta el mejor desempeño y se repite hasta obtener convergencia o un resultado consistente); este algoritmo puede ser integrado en un sistema de cómputo que facilite la clasificación manual de la muestra, y posteriormente, de forma automática, encontrara el nivel óptimo del umbral de clasificación; a continuación, la tabla 5.3.2. muestra los resultados obtenidos tras el desarrollo del análisis.

Tabla 5.3.2. – Tabla que muestra el comportamiento de los índices de eficiencia para diferentes valores de umbral.

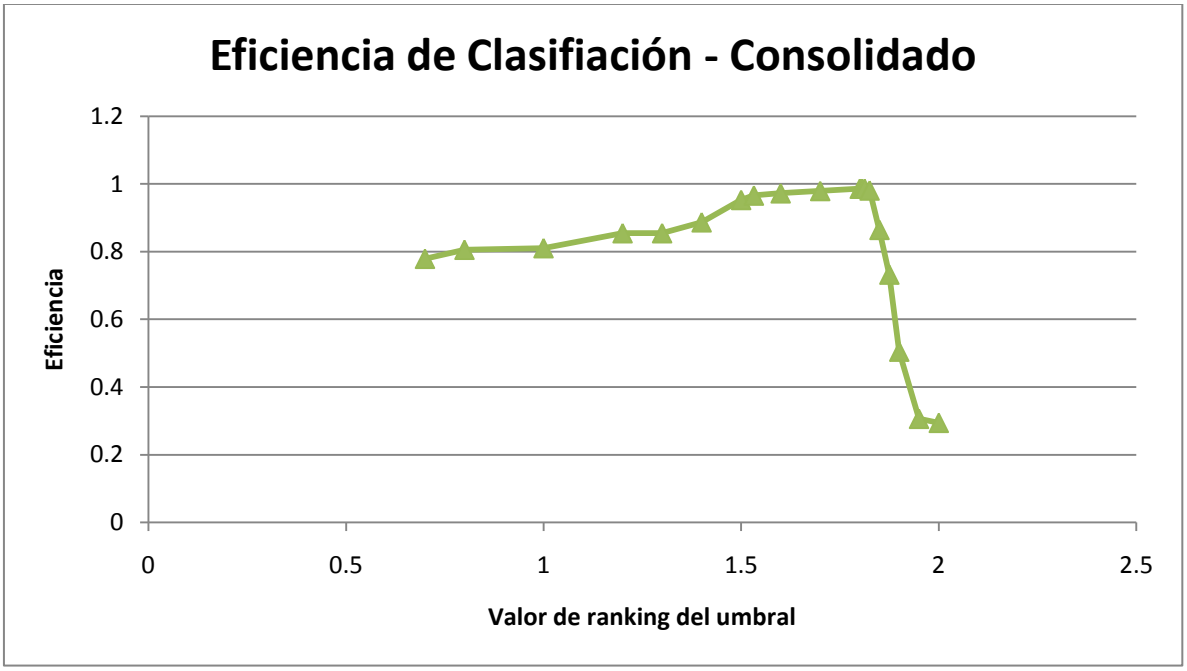
r_{umbral}	$E_{\in}$	$E_{\notin}$
0.7	1	0.779005525
0.8	1	0.805714286
1	1	0.810344828
1.2	1	0.854545455
1.3	1	0.854545455
1.4	1	0.886792453
1.5	1	0.952702703
1.53224907	1	0.965753425
1.6	1	0.972413793
1.7	1	0.979166667
1.8	1	0.986013986
1.80625	1	0.986013986
1.8125	**1**	**0.986013986**
1.825	0.980769231	0.985915493
1.85	0.86440678	0.985185185
1.875	0.732394366	0.991869919
1.9	0.504761905	1
1.95	0.306358382	1
2	0.294444444	1



Gráfica 5.3.1. – Comportamiento de los índices de eficiencia de clasificación con respecto a la variación en el valor de ranking de umbral.

En la tabla 5.3.2. y la gráfica 5.3.1. se puede apreciar un punto de inflexión entre los coeficientes de eficiencia en la clasificación del contenido como perteneciente y no perteneciente en $r_{umbral} = 1.8125$, donde a su vez ambos índices presentan sus valores máximos. Los altos valores en la eficiencia cuando el umbral tiene su valor óptimo se verán afectados de manera directa por características del conjunto de información como pueden ser la diversidad de contenidos y una segmentación menos precisa, lenguajes y estructuras gramaticales regionales, entre otras. Una prueba interesante que queda fuera de los alcances de esta tesis sería incrementar periódicamente el número de conjuntos polares en el conjunto de información e ir evaluando y comparando su desempeño en la clasificación de cada uno de ellos extrapolando los protocolos planteados en el presente documento.

Si se consolida cada pareja de coeficientes de eficiencia por valor de umbral considerando la magnitud más pequeña como la final podemos observar la curva descrita en la gráfica 5.3.2. y representar más claramente la curva de máxima eficiencia en el proceso de clasificación.



Gráfica 5.3.2. – Curva consolidada de eficiencia de clasificación con respecto a la variación en el valor de ranking de umbral.

5.4. Elementos que afectan la eficiencia en el proceso de clasificación

Como pudimos observar en las secciones 5.2. y 5.3. de la presente tesis, el algoritmo ActiveRank es un método viable y de alta eficiencia para la clasificación de contenido en ambientes con conjuntos de información antagónicos, sin embargo, diferentes características pueden afectar de manera significativa su desempeño; a continuación se describen algunas de las más importantes detectadas tras el desarrollo de esta investigación.

1. *Extensión del universo y técnicas de acotamiento de procesos de exploración de la WWW*

Los sistemas de crawling utilizados en la presente tesis pueden ser acotados para discriminar entre páginas pertenecientes a dominios específicos así como limitarse a seguir vínculos hasta cierto nivel de profundidad; esto último, si no está correctamente planteado tiene un impacto directo sobre los procesos subsecuentes, incluyendo la utilización de ActiveRank como plataforma de clasificación automática; un ejemplo sería intentar discriminar páginas cuyo contenido estuviese vinculado estrechamente a una fuente y al mismo tiempo excluir esa fuente en el conjunto de documentos analizados; a pesar de que con ActiveRank se pudiera

clasificar, nunca encontraríamos dicho contenido no por el algoritmo sino por su exclusión en las etapas iniciales de la conformación del universo de información.

2. *Aumento en el número y diversidad de subconjuntos de información*

Relacionado de manera directa al punto anterior, el expandir la cantidad de información recolectada de la WWW trae como consecuencia inmediata la diversificación y el aumento de cúmulos o *clusters* de información en el universo generado; esto dificulta el proceso autoclasificación basado en ActiveRank dado que reduce el intervalo entre el valor de ranking umbral y los valores de ranking medios de los diferentes subconjuntos. Se puede entender como que si se quisieran segmentar dos subconjuntos cuyos valores de ranking medios son muy similares, la eficiencia máxima obtenida tras la clasificación automática sería extremadamente baja. El aumento de subconjuntos de información no afecta el desempeño si el segmento que se desea separar es antagónico a todos los demás, de manera semejante al escenario de estudio en la etapa 1 experimental de la presente tesis, donde el conjunto de páginas de Wikipedia y católicas presentan una alta similitud en términos de ActiveRank entre ellas, pero una gran diferencia con respecto a un tercer conjunto antagónico, en este caso las páginas pornográficas.

3. *Conformación de los vectores de referencia*

Dado que el proceso de autoclasificación basado en ActiveRank utiliza el ranking contra vectores de referencia como criterio de discriminación, afectaciones en la generación de los mismos altera significativamente el desempeño de este proceso. Los principales puntos de falla se encuentran en la incorrecta selección de documentos de referencia para la conformación del vector, en fallas en el proceso de filtrado y depuración del contenido textual del documento, así como en errores sistemáticos producidos por fallas en el parseo, codificación o interpretación de la información digital. Un ejemplo de este tipo de fallas fue lo que nos llevó a desarrollar la extensión de la etapa experimental 2; al operar con un vector de referencia en inglés cuando el universo está conformado por documentos en español inhabilita de manera inmediata la capacidad de discriminación debido a que ninguno de los nodos del vector de referencia es compartido por cualquier otro elemento, o desde otro punto de vista, el vector de referencia se encuentra excluido del universo de información.

4. *Dimensión de los vectores de ActiveRank*

Otro punto fundamental que afecta de manera directa la precisión de ActiveRank es la dimensión de los vectores con los que se trabaja. A pesar de que el algoritmo como tal no limita de manera alguna la extensión o la necesidad de homogenizar el número de componentes de cada vector, existe un punto en el que la expansión

de los vectores no aumenta la eficacia del sistema debido a que su peso relativo es extremadamente pequeño, por lo tanto, es ventajoso económicamente reducir la dimensión de todos los vectores a una longitud óptima, sin embargo, si la dimensión es demasiado pequeña, la información no considerara produce errores importantes en el sistema al otorgar medidas de ranking inconsistentes a la relación o similitud real entre los elementos de la red de información.

5.5. Análisis del dominio “unam.mx”

El análisis extendido de las páginas pertenecientes y relacionadas hasta en tercer grado a aquellas bajo el dominio *unam.mx* se desarrolló con éxito superando la cifra de los 70,000 documentos. Tras corroborar el correcto funcionamiento de todos los sistemas incluyendo el proceso de autoclasificación, la interpretación directa de las gráfica 4.3.10. es que no existe ningún documento con contenido pornográfico en inglés, sin embargo, en la gráfica 4.3.9. podemos observar que hay siete documentos con un valor de ranking inferior a 1.65 con respecto al vector de referencia pornográfico en español, los cuales se revisaron manual e individualmente para ser catalogados; ninguno de ellos pertenece al conjunto de interés, en realidad, son páginas que despliegan errores cuyo contenido es extremadamente escaso, en términos generales solo presentan los encabezados del documento. Tras este análisis más detallado, se encontró una deficiencia en el sistema de crawling la cual permitía que documentos con muy pocos nodos fueran ingresados al sistema ActiveRank, y cualquiera de ellos que compartiera unos cuantos nodos, que adicionalmente correspondían a palabras ambivalentes y no representativas del conjunto, obtendría un valor de ranking cercano a 0. Para corroborar adicionalmente este resultado se hizo una búsqueda manual exhaustiva dentro de la base de datos para intentar detectar direcciones que apuntaran a contenido de éste tipo, sin embargo, de acuerdo a lo esperado, no se encontró ninguna fuente relevante.

No hay contenido pornográfico en el segmento de información procesado en la presente tesis.

Un estudio interesante que queda fuera de los alcances de la presente tesis sería analizar sin acotar el escaneo la WWW partiendo desde diferentes fuentes académicas, institucionales y otras seleccionadas aleatoriamente y observar que tan rápido, en términos de distancia de Hamming, se alcanza una página pornográfica.