Método para el estudio de la producción informativa: medios digitales de Portugal

Method for the analysis of informative production: Digital media in Portugal

Manuel BLÁZQUEZ OCHANDO[1]

Resumen

La presente investigación propone un método de análisis de la producción informativa de los medios de comunicación digitales. El método empleado consiste en el análisis de la producción informativa, recopilada en los canales de sindicación de sus medios de prensa, radio y televisión digital. Se ha tomado como ejemplo de caso, los canales de sindicación de los medios digitales de Portugal, como resultado de la colaboración de la Universidade Nova de Lisboa. Las fuentes de información analizadas, fueron clasificadas según su localización geográfica y temática. El sistema de agregación empleado ha permitido la organización de los medios según su tipología y la obtención de datos tales como el número de informaciones publicadas según regiones, la ratio de noticias por canal de sindicación y región, el número de canales de sindicación y noticias según su cobertura temática, capacidad de recuperación retrospectiva, producción media por día y fracciones horarias y detección de las noticias que mayor impacto tuvieron en la muestra analizada. Los resultados obtenidos han permitido caracterizar y definir los patrones de publicación de los medios de comunicación de Portugal. Por otra parte, se propone un método de clasificación a texto completo que ha permitido localizar y recopilar las informaciones con mayor impacto.

Palabras clave: Agregadores. Extracción de canales de sindicación. Portugal. Producción informativa. Sindicación de contenidos.

Abstract

The present investigation proposes a method of analysis of information production of digital media. The method employed consists of the analysis of information production, which can be collected from syndication channels in the press media, radio and digital television. The syndication channels of digital media in Portugal has been taken as an example case as a result of the collaboration of the Universidade Nova de Lisboa*. The sources of analyzed information were classified according to their geographical location and theme. The aggregation system used has allowed the organization of media according to typology and data obtention, such as: the quantity of information published by regions, news ratio by syndication channel and region, number of syndication channels and news according to thematic coverage, retrospective resilience, average production per day and fractions, and detection of news that had the greatest impact on the sample analyzed. The obtained results have allowed the characterization and definition of the publication patterns of media in Portugal. On the other hand, we proposed a full-text classification method that has allowed locating and retrieving information with the highest impact.*

Keywords*: Aggregators. Extraction of channels of syndication. Portugal. News production. Content syndication.*

[1] Universidad Complutense de Madrid, Facultad de Ciencias de la Documentación, Departamento de Biblioteconomía y Documentación. Calle de la Santísima Trinidad, 37, 28010, Madrid, España. *E-mail:* <manublaz@ucm.es>.

Recebido el 17/2/2016, re-presentado el 31/10/2016 e aceptado para publicación el 31/3/2017.

Introducción

El empleo de canales de sindicación y tecnologías de agregación, viene desempeñando un papel fundamental que aún hoy en día mantiene su relevancia (LIEVROUW, 2015) y se justifica como un método viable para el estudio del derecho a la información, los indicadores sociodemográficos y la producción informativa en torno a la cobertura de los medios de comunicación según apuntan Shumow y Vigon (2016) y Šimec *et al*. (2015). Algunos estudios previos sobre la producción informativa, basada en los canales de sindicación de los medios de comunicación, pueden encontrarse a raíz del desarrollo de la plataforma ReSync presentada por Blázquez Ochando (2013) en el 8º Seminario Hispano-Mexicano de Biblioteconomía y Documentación. Este sistema se concibió para recuperar todas las noticias publicadas por los medios de comunicación de España y México. El investigador podía determinar el periodo de recopilación y clasificar automáticamente los contenidos para su posterior evaluación (BLÁZQUEZ OCHANDO, 2012). Sin embargo, resulta difícil encontrar más estudios que aborden exhaustivamente todos los medios de un país. Es más fácil encontrar estudios parciales en los que se trabaja con una selección de fuentes o canales escogidos. Aunque la literatura científica para este tema en concreto, puede resultar escasa dada la envergadura del planteamiento, sí es posible encontrar otras investigaciones que se especializan en las tecnologías para la revisión y recuperación de canales de sindicación y sus contenidos. En este ámbito cabe destacar el trabajo de Han *et al*. (2009) que presenta un modelo automático de procesamiento de canales de sindicación. También lo es el modelo de extracción de noticias basadas en *WebCrawler* de Adam *et al*. (2010) y el sistema de extracción de citas y temas de noticias de Sarmento *et al*. (2009). En todos los casos se llega a la conclusión de que las mejoras en los sistemas de agregación y recuperación tienen que redundar en métodos que ayuden a filtrar mejor la información. Para este objetivo se está de acuerdo en que es esencial delimitar con precisión las fuentes de información, clasificarlas previamente y mantenerlas actualizadas para evitar la pérdida de contenidos. En parte estos asuntos son los que inspiran la investigación de un método efectivo para recuperar las fuentes de información en un país y tratarlas adecuadamente para estudiar sus contenidos.

En este artículo se plantea el método completo para recuperar los canales de sindicación de Portugal para su posterior procesamiento y análisis en un sistema de agregación de contenidos. El método utilizado consta de los siguientes pasos: (1) Creación de semillas para el *WebCrawler*. (2) Configuración y uso del *WebCrawler*. (3) Selección y extracción de canales de sindicación. (4) Importación y clasificación de canales de sindicación en el agregador. (5) Análisis preliminar de la producción informativa de los medios de comunicación de Portugal, basado en sus canales de sindicación.

Paso 1: Creación de semillas para el *WebCrawler*

Una semilla es un listado de enlaces que tienen en común alguna característica o particularidad específica, de interés para su análisis, por ejemplo, un sector, temática, dominio o contenido homogéneos. Las semillas son empleadas como primer nivel de análisis o punto de partida para los programas *WebCrawler*, analizando y extrayendo los contenidos de cada enlace, para obtener nuevos enlaces, a terceras páginas, dependientes del mismo dominio o sitio *Web*. Por tanto, se necesitan crear listas específicas para el análisis de los medios de comunicación de prensa, radio y televisión de Portugal. Para generar tales listas existen dos métodos: la búsqueda experta en buscadores y la localización de directorios especializados.

- Búsqueda experta en buscadores. El desconocimiento de los principales directorios *Web* especializados, puede motivar una consulta experta en buscadores, con dos objetivos principales: (a) Encontrar directorios o listas de recursos relativos al objeto de estudio; (b) Realizar una consulta directa de los enlaces que se desean localizar. Por ejemplo la consulta "inurl:sapo inurl:\.pt" en buscadores como *Google*, genera una lista de sitios *Web* procedentes del dominio "sapo.pt", uno de los principales portales de noticias.

- Localización de directorios especializados. El conocimiento de directorios *Web* de referencia constituye un excelente punto de partida para obtener los enlaces necesarios de los medios digitales. Por ejemplo, el sitio *Web* "jornais.sapo.pt" contiene una

lista muy completa de los medios de comunicación de prensa en Portugal.

Las semillas de enlaces utilizadas para la extracción de los canales de sindicación de Portugal pueden ser consultadas y descargadas desde las siguientes referencias:

- Semilla de sitios *Web* de prensa de Portugal. (http://mblazquez.es/wp-content/uploads/semilla-prensa-pt.txt)

- Semilla de sitios *Web* de radios digital de Portugal. (http://mblazquez.es/wp-content/uploads/semilla-radio-pt.txt

- Semilla de sitios *Web* de televisión de Portugal. (http://mblazquez.es/wp-content/uploads/semillatelevision-pt.txt).

Paso 2: Configuración y uso del *WebCrawler*

Se ha elegido el *WebCrawler Mbot* para llevar a cabo el análisis de las semillas, quedando disponible para su consulta en la dirección (http://mblazquez.es/mbot/). El motivo principal es su capacidad para extraer y exportar directamente los canales de sindicación de las páginas *Web* de forma automática. Para ello es necesario configurar la profundidad del análisis con un valor próximo a '5 niveles'. También ha de activarse el parámetro 'extracción de canales de sindicación' y la opción 'restricción de sitio'. La profundidad del análisis es el número de niveles de enlazamiento o páginas dependientes. Por ejemplo un sitio *Web* expresado en la semilla, con la designación "p1", puede contener enlaces a otras páginas *Web* del mismo dominio, designadas como "p1.1, p1.2, p1.n" que a su vez pueden contener enlaces a otras páginas designadas como "p1.1.1, p1.1.2, p1.1.n". Cada salto de una página *Web* a otra dentro de un mismo dominio en nivel descendente, se denomina nivel de enlazamiento o nivel de profundidad del análisis. Por otra parte, la opción "restricción de sitio" indica al *WebCrawler* que no tiene que rastrear los enlaces salientes a páginas externas o de dominios diferentes al analizado. De esta forma se evita que analice recursos que nada tienen que ver con el objetivo de los medios de comunicación. Una vez configurado se inicia el proceso de análisis de las semillas.

Paso 3: Selección y extracción de canales de sindicación

Cualquier programa *WebCrawler* puede realizar una extracción de enlaces, limitada a la muestra o semilla de sitios *Web* que se han compilado. Aunque es posible identificar los canales de sindicación por su código HTML *"<link rel= 'alternate' type='application/rss+xml' href='URL-del-canal-de-sindicación'/>"*, no todas las páginas *Web* disponen del método de enlazamiento por defecto. En muchas ocasiones, figuran como enlaces comunes. Ello obliga a que el proceso de extracción de enlaces incluya todas las direcciones URL de la página para su posterior revisión y selección. Esto significa que el resultado de la extracción es una mezcla de enlaces a canales de sindicación y enlaces a contenidos que debe ser cribada antes de continuar todo el proceso.

El cribado de los enlaces se desarrolla atendiendo a cuestiones formales que identifican la presencia de un archivo de tipo RSS o canal de sindicación. Estos aspectos son los siguientes: (a) El tipo de archivo de los enlaces debe ser XML. (b) La primera línea del archivo enlazado debe ser "<?xml version". (c) Debe contener la etiqueta de descripción de canal "<channel>" y la etiqueta de elementos "<item>". Para aplicar estas reglas de verificación, se pueden utilizar expresiones regulares que trabajen sobre el código fuente de todos contenidos vinculados en la lista de enlaces. Las expresiones regulares, también denominadas *Regular Expression* (REGEXP), son notaciones con una sintaxis especial que permite la construcción de patrones de filtrado y búsqueda, cuyas finalidades pueden ser la validación, recuperación de información o datos, o bien el reemplazo de cadenas de caracteres, entre otros usos. Esta tarea suele ser atendida o bien por el propio *WebCrawler*, que dispone de módulos específicos, o bien por el programa de Agregación, que verifica los enlaces antes de proceder a su importación.

Cuando la operación de selección y extracción finaliza, el *WebCrawler* permite la exportación de la lista de enlaces resultante, en formato de texto plano. De esta forma se puede constatar que las direcciones URL corresponden a los de medios de comunicación digitales de Portugal, que pueden ser consultadas y descargadas desde las siguientes referencias:

- Canales de sindicación de prensa de Portugal.

(http://mblazquez.es/wp-content/uploads/feeds-prensa-pt.txt).

- Canales de sindicación de radio de Portugal.

(http://mblazquez.es/wp-content/uploads/feeds-radio-pt.txt).

- Canales de sindicación de televisión de Portugal.

(http://mblazquez.es/wp-content/uploads/feeds-television-pt.txt).

Paso 4: Importación y clasificación de canales de sindicación en el agregador

El siguiente paso consiste en processar los canales de sindicación em um agregador especializado. Em este caso se usará el programa AXYZ (BLÁZQUEZ OCHANDO, 2016a), disponible para su descarga, instalación y experimentación en la siguiente dirección <https://sourceforge.net/projects/axyznews/> (BLÁZQUEZ OCHANDO, 2016b).

La importación de los canales de sindicación se dividide en grupos según el dominio de la fuente original y su categoría temática. Aunque es posible importar listas completas, los canales de sindicación no serían correctamente clasificados y ello es fundamental para poder realizar el estudio de la producción informativa. Por tanto, a cada grupo de canales de sindicación que es importado, se le asignan categorías que definen el tipo de medio (Prensa, Radio, Televisión), la localización geográfica (Según sistema de codificación Nomenclatura das Unidades Territoriais para Fins Estatísticos. II, Norte, Centro, Alentejo, Lisboa, Algarve, Azores y Madeira) y la temática o sección de contenido a la que corresponda. El proceso de importación asistido del agregador AXYZ (BLÁZQUEZ OCHANDO, 2016b), elimina canales de sindicación duplicados y mal codificados de forma automática, obteniendo una lista más reducida que la que se había obtenido en origen. De hecho, la suma de los canales de sindicación de prensa, radio y televisión era de 1.328 referencias que quedaron reducidas a la cifra de 1.023, lo que significa que 305 canales de sindicación no estaban correctamente editados, o bien no fueron validados por problemas en su codificación.

Una vez importados los canales de sindicación en el agregador, es posible ampliar y perfilar en detalle todos los datos de cada uno de ellos. La edición puede realizarse en bloques, lo que facilita el proceso de revisión previo al análisis. Otro aspecto que facilita la localización de problemas en las listas de edición, consiste en el indicador del estado de la clasificación, situado al comienzo de cada fila en el canal de sindicación, que muestra tres puntos verdes, correspondientes a los tres bloques clasificatorios. Cuando uno de los puntos se colorea en rojo, significa que el canal de sindicación no fue clasificado y requiere su revisión, facilitando su rápida identificación visual. Cuando se ha asegurado que todos los canales han sido correctamente clasificados, esto es, que disponen de título, código de idioma, materia y dirección de dominio de la fuente original, entonces puede confirmarse que la lista dispone de los datos mínimos necesarios, para iniciar el análisis de la producción informativa.

Paso 5: Análisis preliminar de las publicaciones de medios de Portugal

El análisis de los canales de Portugal se desarrolló en la Universidade Nova de Lisboa, manteniendo en ejecución el agregador AXYZ (BLÁZQUEZ OCHANDO, 2016b) durante 24 horas desde el día 8 de junio de 2015 a partir de las 16:14:45 horas, obteniendo como resultado, un total de 43.513 noticias de 1.023 fuentes de información. La distribución de los canales de sindicación en los distintos medios de comunicación, demuestra que más de la mitad, corresponden al sector de prensa con 569 enlaces y casi un cuarto a la televisión con 221, que representan el 23% del total. El 65% de los canales recopilados son relativos al área territorial de Portugal (PARLAMENTO EUROPEO, 2016), seguidos de los canales de sindicación de la región del centro y norte de Portugal con un 11% y un 7% respectivamente. Le siguen la región de Alentejo y Madeira con el 4% del total. En suma, los medios de comunicación regional y local alcanzan el 30% de los canales de sindicación disponibles. Los contenidos recuperados en el transcurso del análisis para cada región, demuestran cantidades proporcionales y correlativas al número de canales de que disponen, a excepción de la zona norte de Portugal y Alentejo que llegan a producir casi el doble de noticias que los medios de la zona centro, ya que su *ratio* de noticias es de 11,14 frente a 33,64 y 32,40 (Tabla 1).

Tabla 1. Distribución de canales de sindicación.

Cobertura geográfica	Número de canales de sindicación	Ítems generados en las primeras 24 horas de análisis	*Ratio* de noticias por canal de sindicación
Portugal Azores	20	697	34.85
Portugal Madeira	36	611	16.97
Portugal Lisboa	4	186	46.50
Portugal Algarve	20	306	15.30
Portugal Alentejo	42	1 361	32.40
Portugal Centro	111	1 236	11.14
Portugal Norte	70	2 355	33.64
Portugal País	659	34 670	52.61
África	24	647	26.96
Europa	13	765	58.85
Internacional	24	679	28.29

Fuente: Elaboración propia (2015).

En relación a las categorías más productivas se encuentran Deportes (5.427), Opinión y entrevistas (4.261), Educación, cultura y artes (2.780), Noticias locales (2.780) y Finanzas (2.661). Noticias internacionales, Empleo, Economía y Ciencia quedan relegadas al puesto 13 y sucesivos (Tabla 2). Una mayor producción informativa, implica una mayor demanda de información a nivel social y una prioridad que queda patente en los medios portugueses. Esto significa que el Deporte es más importante que sectores claves para el desarrollo del país, tales como la Economía o la Ciencia y Tecnología. Sin embargo, es muy destacable que la Educación y la Cultura ocupen el tercer lugar de la tabla, lo que indica un apoyo e interés unánime en todos los medios por estos temas.

Tabla 2. Categorías temáticas según canales.

Rango	Categorías temáticas	Número de canales de sindicación	Ítems generados en las primeras 24 horas de análisis
1	Deportes	108	5 427
2	Opinión, entrevistas y crónicas	92	4 261
3	Educación, Cultura y Artes	88	2 780
4	Locales	78	2 776
5	Finanzas y Negocios	42	2 661
6	Noticias multimedia y Podcast	28	2 355
7	Salud	28	1 867
8	Política	39	1 817
9	Primera Página	28	1 806
10	Nacional	46	1 759
11	Miscelánea	34	1 728
12	Sociedad	29	1 643
13	Mundo	35	1 537
14	Empleo	66	1 513
15	Economía	32	1 367
16	Ciencia y Tecnología	21	1 310
17	Ocio	13	920
18	Espiritualidad y Religión	17	877
19	Últimas Noticias y Destacadas	21	842
20	Regionales	35	742
21	Música	19	591
22	Familia	5	408
23	Estilo de vida	10	406
24	Gastronomía	11	392
25	Medio Ambiente	14	327
26	Informática e Internet	26	299
27	Celebridades y Corazón	16	264
28	Cine	11	259
29	Televisión, Series y Programación	11	223
30	Motor	10	188
31	Agenda	3	82

Fuente: Elaboración propia (2015).

No obstante, estos resultados son absolutos, ya que los canales de sindicación almacenan información retrospectiva desde el año 2006, a pesar de que el análisis se produzca en junio de 2015 (Tabla 3). Por tanto, es necesario contrastar estos resultados con los obtenidos durante las 24 horas de ejecución del agregador.

Tabla 3. Recopilación de noticias por años.

Años	Ítems recopilados	Años	*Ítems recopilados*
2001	2	2009	1 495
2002	0	2010	1 194
2003	0	2011	1 337
2004	0	2012	2 331
2005	0	2013	4 106
2006	516	2014	6 483
2007	1 243	2015	18 930
2008	1 547	Suma	39 184

Fuente: Elaboración propia (2015).

El remanente de información es mayor en los últimos años, en concreto a partir del año 2013 y sobre todo en 2014. Esto significa que la información almacenada en los canales de sindicación desaparece con gran rapidez, a razón de 12.000 noticias entre los años 2015 y 2014. Otro dato interesante es que el número de noticias datadas es de 39.184, lo que significa que 4.329 noticias carecen de datación y por tanto no han podido ser evaluadas en las tablas estadísticas. Ello ha sido debido a que el canal de sindicación no proporcionaba tal información, o bien las fechas no estaban normalizadas de acuerdo a los formatos de fecha normalizados ISO 8601 *(Data elements and interchange formats)* o bien RFC 2822 *(Internet Message Format)*. Centrando el análisis en los meses del año 2015, se observa que según se aproxima la fecha en la que fue efectuado el análisis, el número de contenidos recopilados es mayor y ello ratifica una vez más que existe una alta renovación de los contenidos, que desplaza en cuestión de días a los anteriormente publicados.

Si se tiene en cuenta que, en la primera semana del mes de junio, se recupera casi el doble de contenidos que el mes anterior, significa que cada 7 días se pierde el 45% de las noticias publicadas en los canales de sindicación, para dar pie a las nuevas informaciones. La pérdida de contenidos por frecuencia de actualización, viene siendo estudiada por investigadores como Horincar *et al*. (2010). Lograron determinar cuáles era los periodos de recuperación óptimos de los agregadores en las colecciones de canales de sindicación más extensas. Explicaron este hecho mediante una función exponencial cuyo crecimiento aumenta conforme se acerca el día del análisis. Por tanto, la pérdida de información en los canales de sindicación es inversamente proporcional al tiempo que transcurre desde su última publicación, validando la teoría citada anteriormente.

Centrando el análisis en los primeros días del mes de junio de 2015, se observa que el viernes 5 se genera casi el doble de noticias que el resto de días de la primera semana. No obstante, las cifras obtenidas son reducidas en comparación con los días dedicados al análisis, en las que se obtiene un promedio de 2.600 noticias diarias. Esto significa que en un día normal en Portugal los medios de comunicación producen, al menos cerca de 1.500 noticias si se tiene en cuenta la serie de noticias remanente en la semana anterior, con una media de 600 noticias por día (Tabla 4).

Tabla 4. Contenidos de la primera semana de junio de 2015.

Junio de 2015		Ítems recopilados
Día 1 Lunes		308
Día 2 Martes		471
Día 3 Miércoles		530
Día 4 Jueves		649
Día 5 Viernes		1 015
Día 6 Sábado		560
Día 7 Domingo		591
Suma semana anterior		4 124 à Media 589 noticias por día
Día 8 Lunes desde las 16:14:45 h		3.135
Horas	Ítems publicados cada 2 horas	Ítems publicados cada 8 horas
0 - 2h	265	345
2 - 4h	13	
4 - 6h	26	
6 - 8h	81	
8 - 10h	202	1 306
10 - 12h	309	
12 - 14h	291	
14 - 16h	504	
16 - 18h	439	1 484
18 - 20h	378	
20 - 12h	423	
22 - 0h	244	
Día 9 Martes hasta las 16:14:45 h		2.173
Horas	Ítems publicados cada 2 horas	Ítems publicados cada 8 horas
0 - 2h	206	399
2 - 4h	16	
4 - 6h	56	
6 - 8h	121	1 631
8 - 10h	232	
10 - 12h	468	
12 - 14h	436	
14 - 16h	495	
16 - 16:14h	102	
Sin fecha	41	
Suma días 8 y 9		5.308 à Media 2654 noticias por día

Fuente: Elaboración propia (2015).

El desglose horario de la tabla 4 demuestra que las noticias publicadas durante las horas previas al experimento fueron recuperadas completamente y son similares a las obtenidas en la misma franja horaria del día siguiente. Por ejemplo, entre las 00:00 y las 02:00 horas, tanto el día 8 como el día 9 de junio de 2015 se produjeron más de 200 noticias, y en ambos casos se observa una disminución del ritmo productivo en las siguientes 4 horas. También coincide el patrón de máxima cantidad de noticias publicadas en las siguientes 16 horas, como se demuestra entre las 14:00 y las 16:00 horas. Si se comparan estos resultados con los días previos al análisis, se puede deducir que el promedio de actualización de todas las noticias en los canales de sindicación de Portugal es de 24 horas aproximadamente. La marcada variación en las fechas y horas de publicación convalida investigaciones como la de Lee *et al.* (2008). La inclusión de técnicas

adaptativas a los patrones estadísticos de publicación, ayudan a mejorar la eficacia de los procesos de actualización de los canales de sindicación de la colección.

Por otra parte, el análisis detallado de la publicación de noticias según el cuadro horario, puede ayudar a determinar qué eventos y hechos motivan el crecimiento en el número de contenidos publicados. Ello permite estudiar casos que pueden ser de especial interés para los investigadores. Un ejemplo puede encontrarse en la fracción horaria de las 10:00 y las 14:00 horas del día 9 de junio. En comparación con el día anterior, se producen 120 publicaciones más. Estudiando los contenidos obtenidos en esos horarios se descubre que el incremento en el número de publicaciones se debe a la cobertura de la cumbre de países del G7 en vísperas de su celebración en Lisboa.

Además de obtener datos precisos sobre eventos en periodos específicos del cronograma horario, también es posible determinar la temática de los medios de comunicación por días (Tabla 5) y establecer comparaciones e incluso correlaciones con distintos hechos noticiosos que el investigador declare como objeto de estudio.

Tabla 5. Categorías de los contenidos.

Categorías	Ítems publicados el día 8 de Junio de 2015	Ítems publicados el día 9 de Junio de 2015
Primera Página	621	507
Empleo	545	414
Deportes	448	297
Últimas Noticias y destacadas	291	226
Otras categorías	222	6
Sociedad	133	119
Economía	117	95
Regionales	109	74
Mundo	109	80
Nacional	97	50
Finanzas y Negocios	72	62
Opinión, entrevistas y crónicas	52	31
Locales	47	40
Educación, Cultura y Artes	47	50
Política	44	27
Informática e *Internet*	35	17
Noticias Multimedia y Podcast	21	13
Música	21	12
Ciencia y Tecnología	19	5
Salud	16	7
Televisión, series y programación	15	7
Celebridades y corazón	11	9
Espiritualidad y religión	10	8
Estilo de vida	10	3
Cine	8	3
Ocio	7	5
Motor	5	5
Medio ambiente	1	0
Familia	1	1
Gastronomía	1	0
Agenda	0	0
Suma	3 135	2 173

Fuente: Elaboración propia (2015).

https://doi.org/10.1590/2318-08892017000300002

También cabe señalar cómo la producción diaria de contenidos de Educación y Ciencia en los medios de comunicación es muy reducida en comparación con Deporte y Sociedad. Además, ello contrasta con los datos generales obtenidos en la tabla 2, ya que dichas categorías se encontraban mejor posicionadas. También se aprecia cómo el sector empleo se encuentra entre los más productivos en los medios de comunicación. Ello se debe a que existen dos portales de empleo derivados de "sapo.pt". Los portales son "emprego.sapo.pt" y "expressoemprego.pt". Entre ambos se gestionan 61 canales de sindicación especializados en ofertas laborales clasificadas por regiones y ocupaciones, lo que explica su importancia cuantitativa en los resultados.

Discusión y aplicaciones

El método aplicado a los medios de comunicación de Portugal, proporciona datos preliminares para mejorar el conocimiento de la producción informativa. Esto es determinar la temática de la información, su proporción, patrones cronológicos de publicación, grado de obsolescencia de la información en los canales de sindicación, distribución de fuentes y medios según regiones. El método puede ser aplicado al caso de otros países, de los que también pueden extraerse sus fuentes de información, y caracterizar su producción informativa, desde un punto de vista cuantitativo. Ello permitiría el desarrollo de estudios comparativos que podrían determinar la situación de los medios, así como su planificación informativa.

Por otra parte, este método asienta las bases para acometer posteriores investigaciones cualitativas. Esto se debe a que la disposición de los canales de sindicación junto al uso de programas de agregación especializados, permite estudiar cualquier hecho publicado en los medios. Esto es, a través de alertas informativas y filtrado activo de los contenidos. En esta misma línea, ya es posible calcular el impacto de las informaciones que los agregadores recopilan, así como el grado de correlación entre distintos contenidos (BLÁZQUEZ OCHANDO, 2016a). De esta forma se podrá conocer el número de noticias directamente relacionadas con las elegidas por los investigadores, determinar las que resulten similares según un coeficiente de similitud, determinar el impacto de una información durante un periodo concreto y sus consecuencias informativas en otros medios.

También cabe señalar el potencial que supone la extracción y selección masiva de fuentes de información, y en concreto sus canales de sindicación. Independientemente del género, contexto y temática de las mismas, controlar las fuentes a priori, así como su selección y valoración diligente pueden favorecer el desarrollo de herramientas de recuperación de información que reduzcan el ruido de las consultas en determinadas áreas de conocimiento. En otras palabras, las consultas de los usuarios pueden ser dirigidas a un grupo de fuentes seleccionadas, en contraposición a las consultas globales de la *Web*. Ello permite vislumbrar nuevos buscadores especializados que podrían competir en precisión con los actualmente disponibles.

Conclusión

El artículo proporciona un método efectivo para iniciar los trabajos de análisis sobre los canales de sindicación en medios de comunicación digitales en un país. Se plantea la creación de una lista de dominios de los medios de comunicación. Seguidamente se procede a su análisis a través de *WebCrawler* para extraer sus canales de sindicación. A continuación, se utiliza un agregador de contenidos que valide cada canal, desechando aquellos con problemas y errores. Esto da paso a la importación definitiva de los canales y su clasificación previa, a fin de organizar sus principales facetas que identifiquen el tipo de medio, su localización y temática. Finalmente se procesa la información e inicia un estudio cuantitativo en el que se analizan los datos cronológicamente y se profundiza a nivel cualitativo para dar explicación razonada de las cifras y resultados obtenidos.

La investigación aporta las listas de dominios y canales de sindicación de los medios de comunicación digital de Portugal, como resultado del método propuesto. De esta forma la comunidad científica puede descargar estos recursos para elaborar estudios cuantitativos y cualitativos sobre los eventos y noticias de mayor relevancia social y mediática, determinar la evolución y tratamiento de un asunto en los medios

en un intervalo temporal ampliado, incluso realizar estudios comparativos entre los principales grupos de comunicación.

Del estudio cuantitativo se desprenden muchos datos de interés para caracterizar la producción informativa de Portugal. Algunos detalles relevantes son los siguientes: (1) Los medios de comunicación de prensa en Portugal, representan más del 50% del total. Prensa, Radio y Televisión producen en suma más de 1.500 noticias diarias. (2) La alta cadencia de publicación de contenidos en los canales de sindicación, implica que cada 24 horas, el 70% de los contenidos es renovado o actualizado. (3) La hora punta de publicación de noticias en los medios de comunicación de Portugal se sitúa entre las 14:00 y las 16:00 horas.

Las secciones temáticas que más noticias producen son Empleo, Deporte, Sociedad y Economía. Si se tiene en cuenta que la información sobre empleo corresponde a ofertas de trabajo y no se considera como noticia en sí misma, entonces, la información deportiva se sitúa como la más importante. Educación, Cultura, Arte, Ciencia y Tecnologías, en suma, no logran equipararse cuantitativamente al nivel de la sección deportiva, económica o social en el análisis diario.

La capacidad retrospectiva de los canales de sindicación es limitada, a pesar de que 24 695 noticias recuperadas, corresponden al periodo 2006-2014, tal como se ha demostrado con un crecimiento exponencial. Este hecho permite plantear dos reflexiones; la primera relativa a la pérdida de información debida al número finito de noticias con el que se diseñan los canales de sindicación; y la segunda, relativa a cómo perfeccionar los canales de sindicación para que permitan el archivado sistemático de toda la información que producen y su aprovechamiento por los agregadores de contenidos.

Se demuestra que la fracción horaria para el estudio de los eventos publicados por los medios de comunicación es un factor clave para identificar su repercusión social y mediática. También es posible diagnosticar la relevancia de las fuentes en función de la precocidad con que abordan los nuevos eventos y noticias, pudiendo destacar qué medios de comunicación tienen mayor repercusión. Añadido a esto, la repercusión también puede depender del número de canales que los medios destinan a un determinado fin hasta saturar la oferta o a sus competidores. Esto sucede en el caso del tema "empleo" que en su mayor parte está controlado por "sapo.pt".

El método aplicado al caso de Portugal es extrapolable a otros países, usando listas de canales de sindicación de medios, e incluso de otras fuentes, temáticas y contextos diferentes. Ello supone un punto de partida para futuras investigaciones cualitativas en las que poder determinar la influencia, impacto o relevancia de las informaciones, según sus fuentes, así como su grado de correlación, causa, efectos y relaciones, propias del big data informativo.

Referencias

ADAM, G.; BOURAS, C.; POULOPOULOS, V. Efficient extraction of news articles based on RSS crawling. In: MACHINE and web intelligence. *International Conference on*. Piscataway Township, NJ: About IEEE Xplore, 2010. http://dx.doi.org/10.1109/ICMWI.2010.5647851.

BLÁZQUEZ OCHANDO, M. Desarrollo de un sistema de clasificación automática de contenidos en medios de comunicación hispano-mexicanos. In: SEMINARIO HISPANO-MEXICANO DE BIBLIOTECONOMÍA Y DOCUMENTACIÓN, 9., 2012, Madrid. *Anais*... Madrid: e-Lis, 2012. Disponible en: <http://eprints.rclis.org/19031/>. Acceso em: 9 feb. 2016.

BLÁZQUEZ OCHANDO, M. Evaluación del sistema de clasificación automática de contenidos Resync en medios de comunicación españoles y mexicanos. In: Seminario de Hispano-Mexicano de Biblioteconomía y Documentación, 10., 2013, Madrid. *Anais*... Madrid: e-Lis, 2013. Disponible en: <http://eprints.rclis.org/19103/>. Acceso em: 1 feb. 2016.

BLÁZQUEZ OCHANDO, M. Diseño de un agregador para la gestión de los big data informativos. *El profesional de la Información*, v. 25, n. 4, p. 671-683, 2016a. Disponible em: <http://www.elprofesionaldelainformacion.com/contenidos/2016/jul/17.pdf>. Acceso em: 9 ago. 2016.

BLÁZQUEZ OCHANDO, M. *Sistema de vigilancia informacional AXYZnews.* M-005924/2015, 07 sept. 2015, 3 feb. 2016b.

HAN, H.; NORO, T.; TOKUDA, T. An automatic Web news article contents extraction system based on

RSS feeds. *Journal of Web Engineering*, v. 8, n. 3, p. 268, 2009. Available from: <http://dl.acm.org/ citation. cfm? id=2011297>. Cited: Jan. 18, 2016.

HORINCAR, R.; AMANN, B.; ARTIÈRES, T. Best-effort refresh strategies for content-based rss feed aggregation. In: *WEB Information Systems Engineering*. Berlin: Springer, 2010. p. 262-270. http://dx.doi.org/10.1007/978-3-642-17616-6_24

LEE, B. S. *et al*. Design of an RSS Crawler with Adaptive Revisit Manager. In: *SOFTWARE Engineering & Knowledge Engineering*. San Francisco: SEKE, 2008. p. 219-222. Available from: <http://dblp.uni-trier.de/db/conf/seke/seke2008.html>. Cited: Jan. 19, 2016.

LIEVROUW, L. A. Digital media and news. In: *The Routledge Companion to Alternative and Community Media*. London: Routledge, 2015.

PARLAMENTO EUROPEO. La Nomenclatura común de unidades territoriales estadísticas (NUTS). In: *Fichas técnicas sobre la Unión Europea*. Bruxelles: Parlamento Europeo, 2016. Disponible en: <http://www.europarl.europa.eu/atyourservice/es/displayFtu.html>. Acceso en: 9 feb. 2016.

SARMENTO, L.; NUNES, S.; OLIVEIRA, E. Automatic extraction of quotes and topics from news feeds. In: DOCTORAL SYMPOSIUM ON INFORMATICS ENGINEERING, 4., 2009, Portugal. *Anais*… Porto: Faculdade de Engenharia da Universidade do Porto, 2009. Available from: <http://hdl.handle.net/10216/7080>. Cited: Jan. 20, 2016.

SHUMOW, M.; VIGON, M. News diversity and minority audiences: Using real simple syndication (RSS) to assess the democratic functions of Spanish-language media in the digital age. *Journalism Practice*, v. 10, n. 1, p. 52-70, 2016.

ŠIMEC, A. *et al*. D. Using RSS in advertising: Regional trends and global issues. In: INTERNATIONAL CONFERENCE ON ECONOMICS AND BUSINESS ADMINISTRATION (ESA 2015), 2015, Barcelona. Anais… Barcelona: Hrvatska Znanstvena Bibliografija, 2015. (Business and Economics Series, 19). Available from: <http://bib.irb.hr/prikazi-rad?rad=762317>. Cited: Jan. 18, 2016.