EL MANEJO DE DATOS

Aproximación desde los estudios de la información

Georgina Araceli Torres Vargas

Z666.73
L56M3

El manejo de datos. Aproximación desde los estudios de la información / Coordinadora Georgina Araceli Torres Vargas. - México: UNAM. Instituto de Investigaciones Bibliotecológicas y de la Información, 2019.

viii, 116 pp. - Colección: TECNOLOGÍAS DE LA INFORMACIÓN.

ISBN: 978-607-30-2690-1

1. Datos vinculados. 2. Minería de datos. 3. Investigación bibliotecológica.
I. Torres Vargas, Georgina Araceli, coordinadora. II. Ser.

Diseño de portada: Natalia Cristel Gómez Cabral

Primera edición, 2020

Ciudad Universitaria, 04510, Ciudad de México
Impreso y hecho en México
ISBN: 978-607-30-2690-1

Publicación dictaminada
2020

Contenido

Minería de texto aplicada a un diagnóstico de usuarios en Ciencia y Tecnología: aprendizajes para fortalecer la investigación bibliotecológica

Lourdes Feria Basurto
Consultora e investigadora independiente

Introducción

Las dos actividades de divulgación científica a la que menos asisten las familias en México son la *Semana Nacional de la Ciencia* y los *Talleres Itinerantes de Ciencia*, que ofrece el Consejo Nacional de Ciencia y Tecnología junto con sus contrapartes en los estados del país. La edición más reciente de la *Encuesta sobre la Percepción Pública de la Ciencia y la Tecnología* (ENPECYT) (INEGI 2015), preparada por el Instituto Nacional de Estadística y Geografía (INEGI) en conjunto con el Consejo Nacional de Ciencia y Tecnología (Conacyt), identifica como uno de los indicadores de involucramiento en esos temas, por parte de la ciudadanía, el porcentaje de visitas a recintos y actividades vinculadas con la apropiación del conocimiento, e informa que a nivel nacional la *Semana Nacional de Ciencia y Tecnología,* es la que se reporta como la opción menos favorecida en cuanto a asistencia, lo que deja en antepenúltimo y penúltimo sitios las exposiciones tecnológicas

e industriales y los planetarios, que se ven rebasados ampliamente por la asistencia a los cines, parques de diversiones y zoológicos/acuarios.

Ilustración 1: Población que visitó diferentes tipos de recintos

Indicador	2013	2015
Zoológico o acuario	42.2	31.0
Biblioteca pública	24.1	23.0
Museo de ciencia y tecnología	16.3	17.8
Planetario	12.9	12.3
Exposiciones tecnológicas o industriales	18.5	12.6
Semana nacional de ciencia y tecnología	8.2	7.8
Museo (de arte, de cera, natural)	26.4	26.4
Parque de diversión	49.6	38.9
Teatro	22.9	19.1
Cine	NA	55.7

Notas y Llamadas:
/a Población de 18 años y más.
NA No aplica
Los valores no son sumables, dado que se trata de una pregunta de opción múltiple.
Los valores pueden variar debido al redondeo.
Fuente:
INEGI. CONACYT Instituto Nacional de Estadística y Geografía. Consejo Nacional de Ciencia y Tecnología. Encuesta sobre la Percepción Pública de la Ciencia y la Tecnología (ENPECYT) 2013, 2015

Ante esa realidad, en el estado de Colima, el Consejo Estatal de Ciencia y Tecnología (Cecytcol) instrumentó en 2017 un estudio cuyo objetivo era conocer el impacto de la apropiación social de la ciencia en todos los municipios del estado, en las comunidades y en las escuelas de los niveles primaria, secundaria y bachillerato. Para desarrollarlo se partió de una revisión documental y archivística de los últimos tres años fiscales (2014, 2015 y 2016), así como de un levantamiento de datos *in situ* durante los meses de octubre 2017 a abril 2018, a fin de reconocer las áreas de oportunidad

que tienen las actividades de divulgación en la entidad para, en lo futuro, mejorarlas buscando la congruencia con su *Plan estatal de desarrollo* (Colima 2016), que propone la construcción de una economía del conocimiento con mayores oportunidades para los jóvenes, partiendo de una realidad que muestra la persistencia del rezago educativo, una cobertura insuficiente y una baja calidad en la educación, pero con la mirada puesta en impulsar una política a favor de la innovación, el fortalecimiento del vínculo sector productivo - generación de conocimiento, la mejora de la conectividad del estado, así como la reducción de las brechas educacionales y la armonización de la educación con las necesidades del mercado laboral; haciendo énfasis en uno de sus objetivos (II.3.4.1.2) y "ampliar la divulgación de la ciencia y la tecnología en los niveles medio superior y superior" (Colima 2016, 115).

Discursos y narrativas como fuentes de datos

Los insumos informacionales que permitieron obtener testimonios orales en la forma de discursos, historias de vida y narrativas partieron del planteamiento de la pregunta clave que guió el estudio: ¿cómo atraer a más personas a actividades de información y conocimiento?, esto se resolvió estructurando una metodología mixta para el levantamiento de datos, que comprendió seis etapas:

> Etapa 1: Investigación documental y archivística.
> Etapa 2: Observación participante e involucramiento con las comunidades.
> Etapa 3: Etnofotografía y diarios de campo de investigación acción.
> Etapa 4: Encuestas a estudiantes asistentes a los talleres (aplicación de 381 cuestionarios a estudiantes de nivel básico, medio y medio superior).
> Etapa 5: Grupos focales con profesores y con divulgadores de la ciencia.

Etapa 6: Entrevistas a profundidad con profesores y divulgadores de la ciencia.

Para los fines de la presente revisión, se hará énfasis en las etapas 2, 3, 5 y 6 y se describirán a continuación las técnicas aplicadas en cada una de ellas.

Observación participante e involucramiento con las comunidades (Etapa 2)

El trabajo etnográfico comenzó con la observación sistemática y el levantamiento de notas de campo durante catorce semanas en las que se registraron los eventos significativos de cada día junto con las interpretaciones de los informantes. Las observaciones iniciales se centraron en la recopilación de datos generales y abiertos. Este proceso fue importante para recabar antecedentes para una investigación más centrada y también para establecer una buena relación con los informantes, evitar interpretaciones parciales y probar si las preguntas de investigación originales resultaban significativas y pertinentes.

Por otra parte, se realizó una intervención dentro de las actividades de divulgación como asistentes/oyentes entre las personas estudiadas durante un periodo de seis meses, se recopilaron datos mediante la participación continua en los talleres, charlas, etcétera.

Etnofotografía y diarios de campo de investigación-acción (Etapa 3)

Además de las observaciones escritas, los registros y las bitácoras, la investigación cualitativa se apoyó en levantamientos etnofotográficos en imagen fija y video, lo que integró una galería de más de novecientas fotografías y dieciséis videos y audiograbaciones. Como parte de las actividades de investigación-acción dos de los integrantes del grupo de investigación formaron parte activa al integrarse como conferenciantes en la modalidad de "Charla con un

Científico" e impartir en tres diferentes locaciones rurales la conferencia denominada "Los drones y tú", evento que generó el valor agregado de observar una atmósfera de valoración favorable hacia la ciencia y el interés de los asistentes, en su mayoría niños entre los siete y doce años de edad.

Grupos focales y entrevistas a profundidad con profesores y divulgadores de la ciencia (Etapas 5 y 6)

Después de la orientación inicial, se siguió un programa sistemático de entrevistas formales con base en una batería de cuestionamientos relacionados con las preguntas de investigación. Para ello se seleccionaron veintiún informantes clave para investigar los patrones de percepciones. A partir de ese universo, se hicieron catorce entrevistas a profundidad y dos sesiones de grupos focales. La selección de informantes clave se realizó mediante la variante de *muestreo de juicio* cuidando elegir sujetos bien informados, confiables y que pudiesen informar de los datos contextuales y reconocer los elementos significativos así como las interconexiones a medida que se desarrollaban las secuencias de entrevistas. Desde la perspectiva del análisis del impacto se consideraron, en primer lugar, los elementos simbólicos y se registraron observaciones con la debida atención tanto al contexto cultural como a los significados asignados por los involucrados.

Asimismo, con el fin de dar mayor sustento a esta vertiente de la investigación cualitativa, se hizo previamente una revisión cuantitativa de los informes 2014-2016, así como un reporte de talleristas a partir de lo cual se pudieron extraer inferencias validadas estadísticamente.

Metodología para el manejo de datos: experimentación con minería de texto

Las cuatro etapas descritas permitieron recabar fuentes de datos primarias sobre las percepciones ciudadanas, así como los comportamientos y expresiones individuales hacia la divulgación científica. Con ello se produjo una base de conocimiento integrada por documentos fotográficos, informes, entrevistas, conversaciones y las correspondientes notas de trabajo de campo basadas en la observación sistemática realizada durante catorce semanas, cuya evidencia quedó registrada en dieciséis expedientes de transcripciones basadas en audio y videograbaciones a partir de dos grupos focales y catorce entrevistas; un catálogo/bitácora de cerca de mil piezas de fotografía etnográfica catalogadas y analizadas, un diario de campo incluyendo notas de campo semanales y reportes de observación participante, y un archivo digital de cuatrocientos párrafos testimoniales todo lo cual hizo posible identificar unidades de valor para su filtrado y análisis.

Tras el levantamiento de datos cualitativos se trabajó la información mediante minería de texto, técnica que ha sido descrita como

> […] un campo interdisciplinario que combina técnicas de lingüística, computación y estadística para recuperar y extraer información a partir de texto digital (Bergman, Hunter y Rzhetsky 2013, 210); y también como un proceso automatizado para grandes cantidades de datos textuales, no estructurados, para la recuperación, extracción, interpretación y análisis de información (Reilly 2012).

Otros términos con los que se conoce a la minería de texto son: minería de información, arqueología de información, gestión de conocimiento, data mining, etc., dependiendo del autor pero a lo que nos lleva es a que surja "la necesidad de darle un valor adicional a la información documental" (Justicia de la Torre 2017, 2).

Minería de texto tiene que ver con datos textuales no estructurados y el objetivo es que mediante la aplicación de algoritmos de minería informática se transforme la información textual en

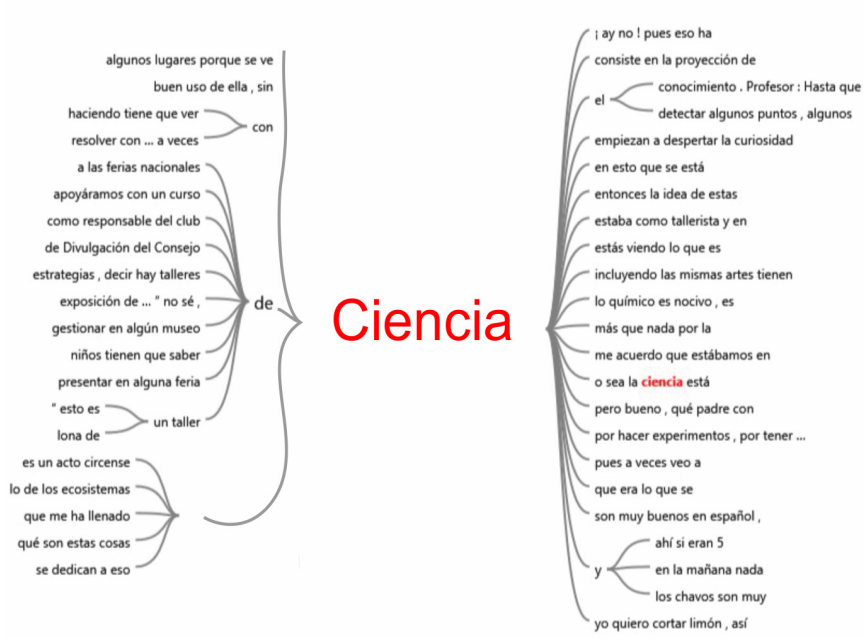
números y pueda identificarse nuevo conocimiento. En síntesis se trata de la aplicación de algoritmos informáticos al texto de las entrevistas; de tal suerte que a partir del lenguaje coloquial no estructurado se generen datos numéricos, vectores e indicadores; lo cual, expresado en términos matemáticos sería: la cuádrupla [DQFR(qi,dj)] donde "D" es un conjunto de vistas lógicas de documentos; "Q" es un conjunto de consultas de usuario; "F"es el marco de trabajo que vamos a usar para modelar y "R(qi,dj)" correspondería a la función *ranking* (Justicia de la Torre 2017).

En el caso particular de los datos cualitativos arrojados, éstos se procesaron mediante un *software* especializado que, en una primera etapa, permitió identificar patrones semánticos a través de los datos para hacer contrastaciones de las distintas historias, para posteriormente estructurar un mapeo semántico que permitió identificar dentro del *corpus* de textos la coocurrencia de las distintas palabras. Para ambos fines se digitalizaron las catorce entrevistas y los dos reportes de grupos focales para integrarlos en la aplicación digital *NVivo 11;* herramienta que permitió, por una parte, entender la frecuencia y asociaciones terminológicas para armar mapas y redes terminológicas; y por la otra, representar cada término en un espacio vectorial donde aquellas palabras con significado similar lograban estar más cerca en el trazado para hacer cálculos de frecuencias y porcentajes. Algunos de ellos se ilustran a continuación.

I*lustración 2:* Porcentaje ponderado de palabras en narrativas

Palabra	Longitud	Conteo	Porcentaje ponderado %
ciencia	5	80	1.16
niños	7	71	1.03
talleres	8	70	1.01
actividades	11	49	0.71
profesor	8	33	0.48
escuelas	8	30	0.43
trabajar	8	28	0.40
taller	6	27	0.39

Ilustración 3: Árboles semánticos



Los resultados de la investigación reflejan que, en general, lo que piensa la comunidad en Colima es que la ciencia es agradable y que cuando se les da a conocer a los niños y a los jóvenes se logra entusiasmarlos sinceramente. Por otra parte, se observa que la difusión sólo permea en las escuelas; que existe articulación entre talleres y programas de estudio; que es necesaria mayor actividad en zonas rurales, y que es necesario sensibilizar a las autoridades y motivar a los padres de familia.

Tabla 1. Tabla de resultados

Porcentaje	Concepto
75%	La Ciencia agradable es posible en talleres.
75%	La difusión sólo permea en escuelas.
75%	Necesaria mayor actividad en zona rural.
69%	Se logra motivar la vocación científica.
44%	Necesario sensibilizar autoridades.
38%	Existe articulación entre talleres y escuela.
31%	Necesario motivar padres de familia.

Resultados

Cabe señalar que previamente a la aplicación del *software* se llevó a cabo un análisis general y una limpieza de datos para proceder al ingreso de la información a fin de producir las tablas porcentuales de palabras más frecuentes, incluyendo la longitud de cada término, el dato sobre la cantidad de ocurrencias en el texto y el porcentaje ponderado de aparición. También se elaboraron tanto conglomerados a partir de palabras, como árboles semánticos que permitieron ver toda la pre y post-coordinación a partir del metadato clave que se eligió como elemento base; con todo ello se pasó a la fase de discernimiento a partir de la observación de vectores de coincidencias.

Conclusiones

Al final del estudio, las conclusiones emanadas se integraron en tres enunciados:

De textos a números y porcentajes. Fue posible a partir de textos no estructurados obtener formas intermedias numéricas que permitieron rescatar y medir aspectos relevantes y de ahí generar

nuevo conocimiento. De haber trabajado en forma manual, no necesariamente se hubieran podido identificar, o el tiempo para lograrlo habría sido mucho mayor.

No todo es inteligencia artificial. Se requiere de la Intervención humana para la limpieza de datos, la integración y la selección de los mismos. Todas las aplicaciones de minería tienen que ver con la participación del investigador y sus colaboradores, quienes hacen posible que el *software* ejecute de manera precisa las funciones necesarias.

Bibliotecas y manejo de datos. Las herramientas, las técnicas, el almacenamiento de datos, la recuperación y los métodos analíticos aún están en proceso de evolución, pero cada vez más las bibliotecas tendrán que fortalecerse en el uso de estos métodos y técnicas para orientar a los investigadores en sus proyectos. Entonces, ¿por qué no se va convirtiendo la biblioteca en el laboratorio natural para la gestión y organización de datos, así como en el área que se haga cargo de la capacitación permanente sobre la alfabetización en datos?

Bibliografía

"Research on tacit knowledge mining of university libraries based on data mining." 13Th International Conference On Service Systems And Service Management (ICSSSM), Service Systems And Service Management (ICSSSM), 2016 13Th International Conference On 1. *IEEE Xplore Digital Library,* 2016.

Botta Ferret E, Cabrera Gato JE. "Minería de textos: una herramienta útil para mejorar la gestión del bibliotecario en el entorno digital". *Acimed* 16, no. 4 (2007).

Bergman, Hunter y Rzhetsky (2013) citado por Dyas-Correia, Sharon, and Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Bernard Reilly (2012) citado por Dyas-Correia, Sharon y Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Cleary P, Garlock K, Novak D, Pullman E, Mann S. "Text Mining 101: What You Should Know. *Serials Librarian.* January 2017;72(1-4):156-159.

Connaway, Lynn y Marie L. Radford. *Research methods in Library and Information Science.* 6a. ed. Santa Barbara, CA, Libraries Unlimited, 2017.

Connaway, Lynn, S., William Harvey, Vanessa Kitzie, y Stephanie Mikitish. *Academic Library Impact: Improving Practice and Essential Areas to Research.* Chicago: Association of College and Research Libraries, OCLC Research, 2017.

Consejo Estatal de Ciencia y Tecnología del Estado de Colima, Consejo Nacional de Ciencia y Tecnología y Gobierno del Estado de Colima. *Estrategia nacional para fomentar y fortalecer la difusión y divulgación de la ciencia, la tecnología y la innovación en las entidades federativas: Colima.* Colima: CECYTCOL, 2014. Trabajo presentado en 21ª Semana Nacional de Ciencia y Tecnología. (Recuperado de: 21SNCT-COLIMA.docxs)

Contreras Barrera, Marcial. Minería de texto: una vision actual. *Bibl. Univ.,* 17, no. 2 (2014), 129-138.

Dyas-Correia, Sharon, and Michelle Alexopoulos. "Text and Data Mining: Searching for Buried Treasures." *Serials Review* 40, no. 3 (September 2014): 210.

Faniel, Ixchel y Lynn S. "Librarians' Perspectives on the Factors Influencing Research Data Management Programs". *College & Research Libraries Journal: 79,* num. 1, (2018).

Instituto Nacional de Estadística y Geografía. Encuesta sobre la percepción pública de la Ciencia y la Tecnología (ENPECYT). México, INEGI, CONACYT, 2015

Justicia de la Torre, María Consuelo. *Nuevas técnicas de minería de textos: aplicaciones.* Granada: Universidad de Granada, 2017.

Mariñelarena-Dondena, Luciana, Marcelo Luis Errecalde y Alejandro Castro Solano. "Extracción de conocimiento con técnicas de minería de textos aplicadas a la psicología" *Revista Argentina de Ciencias del Comportamiento,* 9, no. 2 (2017), 65- 76.

Morris, Walter. "Text Mining for the Social Sciences" *Cornerstone 3 Reports: Interdisciplinary Informatics.* Paper 53 (2011) Santana Mansilla, Pablo; Costaguta, Rossana y Daniela Missio. "Aplicación de algoritmos de clasificación de minería de textos para el reconocimiento de habilidades de e-tutores colaborativos". *Inteligencia Artificial. Revista Iberoamericana de Inteligencia Artificial.* 17, no. 53, (2014), 57-67.

Yu, C. H., Jannasch-Pennell, A., y DiGangi, S. "Compatibility between Text Mining and Qualitative Research in the Perspectives of Grounded Theory, Content Analysis, and Reliability". *The Qualitative Report,* 16, no. 3, (2011), 730-744. http://nsuworks.nova.edu/tqr/vol16/iss3/6

Manejo de datos. Una aproximación desde los estudios de la información. La edición consta de 100 ejemplares. Coordinación editorial, Israel Chávez Reséndiz; revisión especializada, Francisco Xavier González y Ortiz; revisión de pruebas, Valeria Guzmán González; formación editorial, Natalia Cristel Gómez Cabral. Instituto de Investigaciones Bibliotecológicas y de la Información / UNAM. Fue impreso en papel cultural de 90 gr. en los talleres de Grupo Fogra. Año de Juárez 223. Col. Granjas San Antonio. Alcaldía Iztapalapa. Ciudad de México. Se terminó de imprimir en febrero de 2020.