Islas Canarias
Del 15 al 19 de noviembre de 2021

REINGENIERÍA DEL DIRECTORIO DE ACTIVIDADES ECONÓMICAS: DE LOS REGISTROS ADMINISTRATIVOS AL BIG DATA

Amparo Ruiz Gil
EUSTAT
Amparo_ruiz@eustat.eus

Introducción

El Directorio de Actividades Económicas recoge información relativa a los establecimientos ubicados en la Comunidad Autónoma de Euskadi donde se ejerce cualquier de actividad económica, estén sus correspondientes empresas asociadas, ubicadas en la CAE o no.

El Directorio de Empresas constituye la base para cualquier investigación estadística orientada a empresas, por lo que es necesaria una información exhaustiva y veraz. Las necesidades de exhaustividad y calidad de la información han llevado a la utilización de una gran cantidad de recursos económicos y humanos para actualizar anualmente los datos. A pesar de disponer de datos procedentes de registros administrativos prácticamente solo se han utilizado, hasta ahora, como consulta.

Este sistema, aunque ayuda en la codificación de las variables del directorio, no reduce significativamente la carga a las personas informantes, los plazos, ni los costes de la operación, pues es un mero apoyo del personal de campo.

Con el fin de reducir el impacto de la recogida en la sociedad, costes y plazos de disponibilidad de los datos definitivos, EUSTAT se plantea el uso de Registros administrativos y del Big Data para la determinación de algunas variables, permitiendo reducir la carga a las personas informantes, y a las personas encargadas de recoger dichos datos. Incluso se estudia la posibilidad de dar altas y bajas de empresas y establecimientos automáticamente.

Para ello se ha puesto en marcha un proyecto para la obtención de un algoritmo que automatice la actualización del Directorio anualmente, utilizando los diferentes registros administrativos de que se dispone (Impuesto de Actividades Económicas, Seguridad Social, Registros mercantiles, Impuestos tributarios) y obteniendo de ellos información de empleo, actividad, facturación, variables de ubicación, razón social, etc.

Para actualizar ciertos datos, no bien recogidos en los registros, como el teléfono, el mail y la página web de las empresas y establecimientos se plantea el uso de Big Data para obtenerlos mediante web scraping.

Todo esto, manteniendo, por supuesto, los niveles de calidad y fiabilidad altos.

Objetivos

Un Directorio de Actividades o de Empresas consiste en la enumeración individualizada de las empresas, establecimientos, explotaciones u otros entes, donde hay información de su identificación, emplazamiento, descripción de la actividad y uno o más indicadores de tamaño.

Se usa sobre todo como fuente de información para el análisis estadístico de la población de empresas y de su demografía, como instrumento de difusión, para preparar y coordinar las encuestas económicas y, desde luego, para el diseño muestral y la elevación de datos asociados a esas encuestas. Por tanto, es imprescindible disponer de un directorio con un grado de calidad que garantice que las muestras que se extraigan sean las óptimas para la obtención de los datos.

Las necesidades de exhaustividad y calidad de la información que se genera y el hecho de que para su diseño sea necesaria la utilización de datos de diversas fuentes y de distinta naturaleza, hace que anualmente se requieran gran cantidad de recursos económicos y humanos que indaguen, analicen y decidan sobre aquellos datos que serán los que finalmente actualizarán el directorio.

Esta manera de trabajar resulta costosa económicamente, poco ágil, y además repercute en la imagen que el Instituto de Estadística tiene para estas empresas que están recibiendo muchas llamadas consultando por diferentes informaciones de diferentes encuestas.

Ya en el Código de Buenas Prácticas de las Estadísticas Europeas, el Principio 9, habla de '*carga no excesiva para los encuestados*' mencionando, que siempre que sea posible, se utilicen fuentes administrativas. Además, cada vez es más necesario adelantar la obtención de los datos definitivos y su publicación.

En un primer intento de solucionar los problemas de sobrecarga de solicitud de información a las empresas, EUSTAT, elaboró una aplicación llamada Hoja de ruta, desde la que se gestionan todas las peticiones de información a las empresas, además, está ya implantado, de forma generalizada, el uso por parte de las empresas de cuestionarios electrónicos.

No obstante, siguiendo una constante evolución y proceso de mejora continua, desde hace algún tiempo se está trabajando en obtener los datos para actualizar el Directorio de una manera más automatizada, mediante la utilización de la información de registros administrativos (datos de Seguridad Social, de Registros Mercantiles, de Impuestos sobre

Actividades Económicas y de Impuestos tributarios), que hasta hora se utilizaban solo como consulta; y por otro lado, utilizar el Big Data para aprovechar la información que hay en la web sobre las empresas. De esta manera se quiere reducir el coste económico, disminuir los plazos y las molestias ocasionadas a las empresas.

Para llevar a cabo los objetivos propuestos, desde EUSTAT, se han puesto en marcha dos proyectos con unos objetivos concretos:

- Automatización de la actualización del Directorio de Actividades Económicas con información de ficheros administrativos. El objetivo es automatizar el proceso de actualización del directorio de actividades económicas que se realiza con carácter anual.
- Obtención de datos para DIRAE mediante técnicas de Big Data y web scraping:
    - Se persigue la identificación inequívoca de los sitios webs oficiales de las empresas con actividad económica en Euskadi.
    - En fases posteriores se intentará utilizar la información de esas páginas web para actualizar diferentes variables del directorio (teléfonos y mail).

Metodología

Actualización de datos mediante ficheros administrativos

El proyecto comprende el desarrollo de un algoritmo que automatice la actualización del directorio anualmente y los procesos relacionados, en aquellas fases en las que dicha automatización sea posible, estableciendo a su vez mecanismos de validación y seguimiento. En concreto se persigue:

- Actualizar las altas y bajas de las unidades censales que han de formar parte del directorio para el año en cuestión.
- Informar las variables relativas a las unidades de alta como a los cambios que haya habido en unidades ya existentes. Esta actualización tiene que llegar a todas las variables básicas y para todas las unidades que se recogen en el directorio.

Este proyecto consta de las siguientes fases:

- **Inventario de fuentes de datos**. Estudio detallado de las fuentes de información y variables de las que se disponen. Ficheros de Registros administrativos (Seguridad Social, Impuesto de Actividades económicas, Registros Mercantiles, Impuestos tributarios). Otro tipo de ficheros con información de empresas (Infocid).
- **Definición de la parcelación o segmentación de las entidades**. Las empresas y establecimientos de distinta naturaleza que presentan características que les son propias y difieren del resto.

Se ha hecho un primer diseño de esa parcelación:

- Empresas por tipología:
  - Autónomos
  - Autónomos con personal a cargo
  - Cooperativas
  - Comunidades de bienes y Sociedades civiles
  - Administraciones públicas
  - Entidades financieras
  - Microempresas
  - Empresas de más de 20 trabajadores
  - Otras casuísticas

- Ubicación:
  - Con sede en el estado
  - Con sede en el País Vasco

- Tamaño de la empresa:
  - Empresas con una sola UL.
  - Empresas con más de una UL.

- Tipo de actividad:
  - Comercios
  - Mercadillos
  - Profesiones liberales
  - Empresas del sector agrícola

- **Análisis del flujo de los procesos de ejecución de la operación.** La operación del DIRAE comprende distintos procesos:
  - Recogida de la información según diferentes encuestas económicas y operaciones específicas para detectar altas, bajas y modificaciones.
  - Codificación y grabación.
  - Depuración y validación de los datos.
  - Actualización del Directorio anual.
  - Explotación de los datos.
  - Difusión

- **Análisis exploratorio de los datos**, que tiene por objeto conseguir un entendimiento básico de los mismos y de las relaciones que se establecen entre ellos. Este proceso permitirá a su vez determinar qué tipo de errores o inconsistencias deberán ser modificados y/o eliminados en la siguiente fase del proceso.

- **Limpieza de datos o data cleansing**:
    - Identificación de datos incompletos, incorrectos, y/o no pertinentes, como duplicidades y outliers.
    - La modificación y/o eliminación de estos "datos sucios".
    - Filtrar unidades que no interesan.
    - Identificación de mismas unidades. Una de las cuestiones que genera más problemas en todo el proceso es la identificación de las mismas unidades en fuentes de datos diferentes o incluso, en las mismas fuentes.

- **Transformación de los datos**:
    - La armonización de unidades y conceptos entre datos de distintas fuentes. Este proceso implica la conversión, codificación y normalización de las variables (direcciones, nombres, etc.).
    - La generación de otro tipo de variables auxiliares necesarias para abordar el proceso.

- **Fusión de registros de distintas fuentes**.

- **Análisis del flujo de empresas**: en esta fase se aborda la identificación de unidades que son altas o bajas de UJ, UL o UAL en relación a las que ya existen en el directorio, y la identificación de las modificaciones, unidades que ya existen en el Directorio que han podido tener algún cambio en alguna de sus variables.

  <u>Identificación de Altas</u>

  Las altas pueden ser de varios tipos:

  - Altas "puras": nuevas empresas, con UJ, UL y UA nuevas en el directorio, donde nunca antes ha figurado ni el CIF ni el establecimiento.
  - Alta de un establecimiento de una UJ que ya figura en el DIRAE.
  - Altas de CIF que no figuran en el directorio, y por tanto es un alta de UJ, pero su Unidad Local, está ya en el Directorio bajo otro CIF. Habría que comprobar que:
    - la empresa bajo la que estaba el establecimiento no tiene otros o han pasado a otro CIF => darla de baja
    - que tenga varios establecimientos y se queda con alguno => modificación

  <u>Identificación de Bajas</u>

  Ahora mismo ya se hace un proceso automático para dar bajas en el Directorio, pero es muy restrictivo. Con este nuevo proceso se quiere ampliar a más casos.

- **Validación y depuración**. Esta fase tiene por objeto la identificación de posibles errores o incoherencias con el resto de la información que contiene el directorio, así como el diseño de mecanismos de control y seguimiento que permitan garantizar la calidad de los resultados obtenidos en el proceso.

- **Generación de los outputs del proceso:** generación de las tablas con los movimientos definitivos para actualizar el Directorio.

Problemas a los que hacer frente

- ✓ **Armonización y normalización de las variables:**

Diferencias en la forma de recoger la información en cada fichero administrativo, lo que genera muchos problemas a la hora de identificar empresas y unidades locales, a la hora de poder hacer cruces o de poder introducir esa información con los formatos necesarios en las base de datos del Directorio.

- ✓ **Identificación de las unidades (tanto unidades jurídicas como unidades locales):**

1. *Identificación de Unidades Jurídicas. Cambio CIF:*

Análisis de casuísticas:

- Dentro de una misma fuente de datos
- En distintas fuentes de datos

2. *Identificación de Unidades Locales:*

Identificación de una Unidad Local que permanece aunque cambia su titularidad jurídica. La identificación de estas unidades de manera automatizada puede ser complicada.

Criterios para determinar que estamos ante una misma UL:

- Misma ubicación física (misma dirección).Aquí nos encontramos con el problema de cómo están recogidas las direcciones, no están normalizadas en muchos casos.
- Misma CNAE.
- Nombre local: este criterio tiene escasa aplicación porque en la mayoría de los ficheros administrativos no aparece la denominación del establecimiento, sino el de la razón social.
- Relación entre propietarios, para identificar aquellas situaciones en las que algunas de las personas socias de la compañía se "queda" con el negocio.
- Misma estructura de empresa: parecido número de trabajadores, parecida CNAE, parecido tamaño del establecimiento. Se comparan incluso los m^2 de las licencias…

Parece claro que habrá casuísticas complejas que habrá que investigar en campo.

Actualización de datos mediante Big Data

Se llama Big Data a un conjunto de datos masivos, que tienen una gran variedad y se producen a gran velocidad. Suelen ser datos no estructurados y provienen de fuentes muy diversas. Estas fuentes son entre otras: sensores (en carreteras), teléfonos móviles, mensajes de redes sociales, páginas web, información de satélites, información de transacciones financieras y comerciales por las tarjetas de crédito, etc.

Todo esto abre nuevas oportunidades a la estadística, ya que por ejemplo se pueden obtener datos con más frecuencia, se pueden describir los fenómenos antes de tener las encuestas y, por lo tanto, adelantar las publicaciones de datos; se pueden reducir costes, se puede acceder a nuevas fuentes de información, y en algunos casos, se podría dejar de hacer la encuestación por los métodos habituales.

El uso de estos datos para la estadística oficial es muy reciente, Naciones Unidades empieza con el Big Data en 2014, el ESS (European Statistical System), también empieza en esta época desarrollando un Plan de acción y una hoja de ruta.

El Big Data tiene sus limitaciones y tiene que ver sobre todo con la calidad de los datos:

- Difícil acceso a los datos
- Usabilidad de los datos: estabilidad en el tiempo
- Dificultades metodológicas: selección, representatividad y validez
- Problemas con la privacidad y la confidencialidad de los datos
- Infraestructura y seguridad
- Necesidad de nuevas destrezas
- Dificultad en el paso de la investigación a la producción.

A pesar de esas limitaciones, el Big Data, es una apuesta clara para la estadística oficial. Así lo menciona ya el documento de la *Visión 2020 del ESS,* remarcando que esta nueva fuente de datos ofrece una gran oportunidad para mejorar la puntualidad y relevancia de las estadísticas oficiales, así como la reducción de la carga de trabajo a los informantes. EUSTAT ha aceptado este nuevo reto y ya en 2017 comenzó el primer proyecto con Big Data para obtener los precios de los hoteles de la C.A. de Euskadi mediante técnicas de web scraping, analizando diversas páginas web que ofrecen esta información, en concreto, plataformas de reserva online de habitaciones (Booking y Expedia).

Otro de los proyectos de Big Data en el que se ha embarcado EUSTAT ha sido para el Directorio de Actividades Económicas de Euskadi, en este caso, se persigue la identificación inequívoca de los sitios webs oficiales de las empresas con actividad económica en Euskadi. En fases posteriores se intentará utilizar la información de esas páginas web para actualizar diferentes variables del directorio.

La primera fase del proyecto se fundamenta en el proyecto ESSnet Big Data de Eurostat (llamado SummaIstat Scraping Tools), que proporciona una serie de herramientas open source. El objetivo es localizar y actualizar información de las Unidades Jurídicas del Directorio de Actividades Económicas de Euskadi (DIRAE) en base a técnicas de web scraping/crawling y modelos deterministas para calcular la URL oficial de cada Unidad Jurídica. El proceso incluye la utilización de un motor de búsqueda como principal fuente de información. Se inició el estudio con 2.400 empresas (una muestra del DIRAE entre las que ya tienen la URL en el Directorio). El siguiente paso ha sido ejecutar el proceso con lotes de empresas 3.000, ya sin conocer su URL de antemano, y ofrecer esa información a DIRAE.

En este proyecto se hace uso de varias herramientas creadas en el desarrollo del proyecto europeo ESSnet BIG DATA:

- URLSEARCHER: permite realizar búsquedas mediante BING y GOOGLE. Toma como entrada una lista de nombres y, para cada uno de ellos, lanza una consulta contra el motor de búsqueda seleccionado mediante web scraping. Posteriormente, la herramienta escribe el conteniendo de las primeras 10 URL devueltas por el motor de búsqueda. Se ha incluido una blacklist para evitar la captura de URL no interesantes (Infocif, Empresalia, etc.).

- ROOTJUICE: recopila y almacena datos de la web. Recibe como input un listado de URL y, en base a cierta parametrización, recopila y almacena datos de las webs listadas. Descarga el contenido de sitios web.

- URLSCORER: asigna una puntuación de las páginas según su contenido. Para cada empresa analizada, asigna una puntuación a cada URL candidata a ser el sitio web oficial en función de la presencia o no de ciertos parámetros; cada parámetro tiene un peso asociado. El proceso de URLScorer finaliza con la generación de un fichero que permite (como entrada de datos pero, también, como entrenamiento) a un modelo de Machine Learning determinar si una URL es la mejor candidata para ser el sitio web oficial de una empresa en cuestión.

- URLMATCHTABLEGENERATOR: se trata de un programa Java que combina, por una parte, el archivo obtenido como resultado del programa UrlScorer, con una lista de sitios web de empresas correctas conocidos de antemano. Sirve para generar el conjunto de entrenamiento para que el modelo sea capaz de reconocer la URL correcta para cada empresa; sin conocer de antemano el sitio web oficial. Aquí se utilizan técnicas de Machine learning.

- PIPELINE: configura todo el proceso y parámetros.

- CNAE-EXTRACTOR: herramienta que identifica la actividad económica (basado en diccionarios de A38) en cada URL obtenida. En el motor de búsqueda Bing se incluye una búsqueda por contenidos (nombre de empresa, ubicación…). En este contenido descargado de cada web se utilizarán técnicas de Procesamiento

de Lenguaje Natural para inferir la actividad económica de la empresa; en caso de encontrarse y coincidir con el que sabemos es el correcto, se le asigna una puntuación que servirá, junto con el resto de parámetros del modelo determinista, para deducir la probabilidad de que sea el sitio web oficial.

- DOMAIN SCORER: agrupa las URL descargadas por dominio para calcular un vector resumen de puntuación.
- DICCIONARIOS DE ACTIVIDADES ECONÓMICAS: A partir del procesamiento del lenguaje natural y la lematización de sustantivos, se crean diccionarios a la agregación A38 de las CNAE. El *Procesamiento del Lenguaje Natural* (NLP por sus siglas en inglés) consiste, básicamente, en el desarrollo de aplicaciones y servicios capaces de interpretar los lenguajes humanos. En NLP se analiza el leguaje humano, se interpreta y se da significado para que pueda ser utilizado de manera práctica. En nuestro caso, a fin de ser capaces de extraer la CNAE de una empresa, se extrajeron del directorio de empresas una serie de URL para, en base a su contenido web, generar unos diccionarios de terminología por ámbito laboral. El primer paso consistió en clasificar las webs del directorio de empresas en base a su CNAE para poder escoger después un subgrupo de ellas sobre las que trabajar en la creación de diccionarios.

Problemas a los que hacer frente

- Problemática en la descarga de las páginas. Tecnología evolutiva que dificulta el scrapeo de las páginas. Se ha modificado el scrapeo para capturar más paginas (javascript, carga retardada de contenidos,…).
- Evitar la captura de un dominio en más de una ocasión

Desarrollos en marcha:

- Detección de idiomas de las páginas (incluido euskera).
- Detección de ofertas de trabajo en las webs de las empresas.
- Detección de tienda online en las webs y su plataforma.
- Detección de Extranet-s / Áreas de clientes en las webs.
- Mortalidad de las empresas: Hacer uso de Google para obtener listados de empresas que probablemente estén de baja.
- Detección de URL que tenemos en DIRAE y son obsoletas.
- Análisis de procedimientos para conseguir información de las empresas pequeñas (las de 1, 2 y 3 de empleo). Estas no suelen tener una página web

propia, sino que utilizan Facebook, Instagram… Explorar la descarga de la página principal de Facebook, sobre todo para empresas "pequeñas".

Desarrollos futuros:

- Aplicar métodos de NLP para la extracción de características de las empresas.
- Aplicar redes neuronales para determinar la clasificación CNAE.

Resultados

Actualización de datos mediante ficheros administrativos

Los últimos datos de actualización directa del DIRAE mediante ficheros administrativos son de 2019. Se actualizan exclusivamente bajas y modificaciones, no altas. En concreto se actualizaron con este sistema 41. 480 (24,2%):

- Bajas: 4.772 (2,8%)
- Modificaciones: 36.708 (21,4%)

Este número de registros actualizados actualmente se espera aumentar con el nuevo algoritmo con el que se está trabajando, además, no limitándose solo a bajas y modificciones, actualizando también altas.

Actualización de datos mediante Big Data

Por un lado, se ha testado la herramienta con 2.400 webs que teníamos en el DIRAE pertenecientes a 2.400 empresas respectivamente. Este testeo ha dado como resultado que más del 85% de las webs eran buenas.

Por otro, se ha lanzado otro paquete de unas 3.000 empresas que no tienen página web y se han recuperado un 50% de webs correctas. Aquí hay que especificar, que en muchos casos, no se recupera ninguna web porque las empresas no la tienen, dato que nos ha resultado llamativo. Esto último ocurre principalmente en los casos de los autónomos.

También hay que decir, que una cosa es encontrar la web correcta de una empresa y otra es poder descargar esa página y acceder a sus datos. De las webs encontradas, no de todas se ha podido descargar la página debido a que son web dinámicas o que utilizan javascript para generar contenidos, entre otras casuísticas.

De cada empresa obtenemos un mínimo de 10 URLs y 20 páginas máximo de cada URL. Lo que hace que la cantidad de información a almacenar sea muy grande.

Conclusión

Los registros administrativos y el Big Data son ya unas herramientas indispensables para la estadística. En el caso del Directorio de Actividades Económicas pueden ser herramientas fundamentales a la hora de afrontar el futuro. Su utilización va a tener consecuencias importantes en:

- ✓ Reducción de la carga de trabajo de los informantes
- ✓ Reducción de los plazos de publicación
- ✓ Reducción de los costes económicos

En general, la utilización de nuevas fuentes de datos en la estadística oficial es fundamental si no quiere quedarse atrás y perder el protagonismo en la difusión de datos ante la proliferación de estadísticas producidas por otras organizaciones y empresas que ya las están utilizando.

Está claro, que para llevar a cabo este nuevo reto, habrá que afrontar cambios en la formación, la metodología, la infraestructura y la tecnología utilizada hasta ahora. Seguramente esta nueva forma de trabajo no va a sustituir por completo a la tradicional mediante encuestas, pero será un complemento fundamental.

Principales referencias bibliográficas

Rglamento (CE) No 223/2009 del Parlamento Europeo y del Consejo de 11 de marzo de 2009.
Sistema Estadístico Europeo (2011). Código de Buenas Prácticas de las Estadísticas Europeas.
Eurostat (2016). ESS Visión 2020: Building the future of European statistics.
ESSnet Big Data II. Grant Agreement Number: 847375-2018-NL-BIG DATA.