Proyecto piloto sobre viabilidad de usar Internet como fuente de datos (segunda edición)

Destacados

El ONTSI publica los resultados de la segunda edición del proyecto piloto sobre la viabilidad de utilizar Internet como fuente de datos (IaD). IaD hace referencia al uso de técnicas avanzadas de análisis de datos que pueden servirse de Internet como fuente complementaria o sustitutiva de fuentes tradicionales de datos estadísticos. La utilización de este tipo de técnicas permite el ahorro de recursos y tiempo respecto de técnicas tradicionales, y están caracterizadas por su automatismo, su carácter no intrusivo, al no requerir la participación activa de terceros, y su exhaustividad, ya que permite la exploración completa del universo analizado.

En esta edición se ha realizado el análisis de las siguientes cuestiones:

1. Análisis de comercio electrónico en páginas web de empresas españolas.
2. Análisis de la oferta y demanda de perfiles profesionales de las Tecnologías de la Información, las Comunicaciones y los Contenidos (TICC) y en España en portales de empleo, sitios webs de empresas del sector TICC, webs oficiales relativas a títulos universitarios, de formación profesional y cursos de formación en empresas.

En ambos casos se han utilizado técnicas avanzadas de clasificación automática, análisis de datos y aprendizaje automático para detectar y caracterizar tanto el comercio electrónico como los perfiles profesionales. La automatización pretende evitar o minimizar las tareas de exploración o anotación manual de sitios web.

Se ha desarrollado una herramienta de visualización que permite hacer una completa explotación de los resultados. La herramienta de visualización está disponible en los siguientes enlaces:

- Comercio electrónico: http://iad.ontsi.es/B2C/
- Perfiles profesionales: http://iad.ontsi.es/perfilado/
- Perfiles profesionales según la clasificación nacional de ocupaciones (CNO): http://iad.ontsi.es/perfiles_CNO/

A continuación se resumen los resultados obtenidos en esta segunda edición.

- Para realizar el estudio, se han analizado la información de 742.484 empresas procedente del Registro Mercantil de los siguientes sector de actividad:

CNAE	Denominación
10 – 18	Alimentación, textil y otros
19 – 23	Coquerías, plásticos y otros
24 – 25	Metalurgia y productos metálicos
26 – 33	Productos informáticos, electrónicos y mecánicos
35 – 39	Energía y agua
41 – 43	Construcción
45 – 47	Venta y reparación de vehículos
49 – 53	Transporte y almacenamiento
55	Servicios de alojamiento
58 – 63	Información y comunicaciones
68	Actividades inmobiliarias
69 – 74	Actividades profesionales, científicas y técnicas
77 – 82	Actividades administrativas y servicios auxiliares

- Estas empresas representan aproximadamente el 84% del total de empresas registradas en el Registro Mercantil.
- Se parte de un conjunto de 724.848 empresas, de las cuales 202.450 disponían de página web. Después de depurar las urls obtenidas, se analizaron 171.027 dominios (23%).
- En mayo de 2016, el 13% de las empresas españolas ofrecían servicios de comercio electrónico en su página web.
- Las empresas de las ramas de actividad de servicios de alojamiento son las que mayor oferta de comercio electrónico presentan, alcanzando el 76% del total de empresas.
- El siguiente grupo en importancia es el de las empresas de venta y reparación de vehículos, con un 20%. Esta posición se explica por el peso del sector de las empresas de comercio al por menor en este grupo.
- Le sigue la rama de actividad de información y comunicaciones (14%), las de alimentación, textil y otros (14%), las de actividades administrativas y servicios auxiliares (13%) y las actividades inmobiliarias (10%).
- Atendiendo al tamaño de empresa, las que más comercio electrónico ofrecen en su página web son las de grandes empresas de más de 250 empleados, casi el 16% de estas empresas ofrecen estos servicios. Destacan también las PYMES de 50 a 249 empleados, el 15% de estas tienen comercio electrónico. En el caso de las PYMES de 10 a 49 empleados el porcentaje es del 13%, y para la microempresas del 11,5%.
- Por subsectores, destaca que el 90% de las grandes empresas y de las PYMES de 50 a 249 empleados de las ramas de actividad de servicios de alojamiento ofrecen servicios de comercio electrónico en su página web.
- Se ha analizado también la presencia de determinados términos desagregados respecto a la presencia de B2C. Las alusiones a redes sociales

como “Facebook”, “Twitter” o simplemente “email” son muy superiores en empresas con B2C.

- También son significativas las diferencias en lo que respecto al pago y envío. Además de términos como pago, pedido, domicilio, descuento, aparecen en las páginas web con B2C términos como paypal, mastercard, visa, tarjeta.
- Se ha realizado también un análisis de la diversidad de idiomas en las páginas analizadas. El español es el idioma dominante en el 74% de las webs, seguido del inglés dominante en el 15%, tendencia que se repite en todos los sectores CNAE. Sí aparecen algunas diferencias en los terceros y sucesivos idiomas. Por ejemplo, en servicios de alojamiento, el tercer idioma dominante es el catalán, mientras que en coquerías y plásticos, el tercer idioma parece ser el francés. Asimismo, en el sector de alojamiento e inmobiliario, la presencia de idiomas diferentes al castellano es sensiblemente mayor para sitios que tienen B2C, lo cual es lógico, ya que se pretende alcanzar la mayor población posible de potenciales clientes.
- En el análisis geográfico por municipios, se constata que Madrid y Barcelona concentran la mayor parte del B2C. Le siguen Valencia, Palma de Mallorca, Málaga, Zaragoza y Sevilla.
- Por provincias, Madrid y Barcelona son las que totalizan mayor número de empresas con B2C, 4.106 y 3.692 respectivamente. Le sigue Valencia, Alicante y Baleares con 1.214, 1.082 y 1.079 empresas respectivamente.
- Por CCAA, Cataluña, Madrid y Andalucía son las que más empresas tienen B2C. Le sigue la Comunidad Valenciana, Galicia y Castilla y León. Un tercer grupo lo conforman País Vasco, Aragón, Canarias y Castilla La Mancha.

- El análisis de datos proporciona una descripción de los perfiles de empleo más demandados mediante listas de términos relevantes.
- El perfilado automático es una herramienta de interés de cara a realizar un análisis y seguimiento de la oferta de empleo, y del empleo TIC en particular.
- La herramienta de extracción automática de perfiles detecta especializaciones de interés actual en el sector TIC. Los principales perfiles son:

Perfil profesional	%
Programador java, programador para la web	12,0
Experto en marketing, especializado en ventas	11,3
Experto en marketing, en sus facetas de comunicación y explotación de redes sociales	8,0
Técnico de soporte de sistemas	8,7
Administrador de sistemas y bases de datos	7,6
Analista programador con especialización en banca y seguros	8,4
Experto en seguridad	6,3
Analista de datos	5,0
Diseñador web	4,9
Técnico de redes de telecomunicación	4,1
Técnico de redes de datos	3,8

Otros	19,9
Total	100

- No obstante se detectan limitaciones intrínsecas al empleo de herramientas de perfilados completamente automáticas, que se pueden limitar con una reducida supervisión humana.

3.1 Perfiles de empleo en sitios web de empresas TIC

- Los resultados obtenidos en el análisis de las ofertas de empleo en los sitios web de empresas TIC muestran que los perfiles más relevantes coinciden con los obtenidos en el análisis de ofertas de empleo en portales de empleo, lo que confirma la relevancia de dichos perfiles.
- Destacan la importancia de perfiles de gran actualidad como es el relacionado con expertos en big data, o los relacionados con el desarrollo de aplicaciones móviles.

3.2 Perfilado jerárquico en portales de empleo basado en la Clasificación nacional de Ocupaciones

- El estudio incluye un método alternativo para la categorización jerárquica de las ofertas de trabajo, que se apoya en la taxonomía de la Clasificación Nacional de Ocupaciones (CNO).
- Se ha desarrollado un algoritmo de análisis de las ofertas de empleo publicadas en mayo de 2016 en uno de los portales de empleo.
- El algoritmo asigna automáticamente cada oferta de empleo a un número máximo de tres CNOs diferentes. Se tiene en cuenta, por tanto, que cada oferta de empleo puede tener componentes asociadas a varias categorías del CNO.
- Las diez principales ocupaciones encontradas son las siguientes:
    1. La ocupación más demanda es la de analistas, programadores y diseñadores Web y multimedia, apareciendo en el 32,4% de las ofertas de empleo. El perfil más importante dentro de esta ocupación está caracterizado por los términos java, javascript, spring, analista y equipo.
    2. Le sigue en importancia la ocupación de programadores informáticos, que se corresponde con el 26,2% de las ofertas. Las habilidades más demandas configuran un perfil caracterizados por los términos java, j2ee, spring, aplicaciones, sql y analista.
    3. La tercera ocupación por orden se corresponde con los profesionales de la venta de TIC (25,6% de las ofertas). El perfil más relevante se obtiene de la agrupación de términos como comercial, clientes, comunicación, gestión e inglés.
    4. En cuarta posición se encuentra la ocupación de técnicos de ingeniería de las telecomunicaciones, (16% de las ofertas). Las características más relevantes del principal perdil son técnicos, gestión, clientes, soporte, equipo e inglés.
    5. Los técnicos en operaciones de sistemas informáticos son la quinta ocupación en importancia, aglutinando el 14,3% de las ofertas. El

principal perfil requiere habilidades relacionadas con los términos java, j2ee, sping, sql, hibernate y clientes

6. El 12,4% de las ofertas se corresponde con la ocupación de diseñadores gráficos y multimedia, estando caracterizado el principal perfil por los términos marketing, publicidad, comunicación, digital, redes_sociales.
7. La séptima ocupación por orden de importancia corresponde a los técnicos en asistencia al usuario de tecnologías de la información, aglutinando el 10,7% de las ofertas de empleo. El perfil más relevante demanda habilidades relacionadas con clientes, soporte, incidencias, gestión, equipo e inglés.
8. Los analistas y diseñadores de software y multimedia se encuentran en la octava posición, suponiendo el 10,4% de las ofertas de empleo. Los términos java, j2ee, spring, web, software y Oracle caracterizan las habilidades del principal perfil demandado.
9. En novena posición se sitúa la ocupación de analista de sistema, con el 9,7% de las ofertas. El principal perfil viene determinado por los términos java, j2ee, spring, web, software y Oracle.
10. La décima posición corresponde a la de los diseñadores y administradores de bases de datos, presente en el 7,55% de las ofertas. Los términos que caracterizan al principal perfil son equipo, Linux, aplicaciones, clientes, servicios y servidores.

- En el análisis geográfico por provincias, Madrid es la provincia con mayor número de ofertas de empleo, seguida por Barcelona. En mayo de 2016 se encontraron 1.485 y 927 ofertas respectivamente.
- Las provincias limítrofes con Madrid muestran un número muy reducido de ofertas de trabajo TICC.

3.3 Análisis de programas formativos TICC

- El objetivo de esta parte del proyecto es la caracterización de la oferta formativa a partir de la información capturada en sitios web oficiales relativos a títulos universitarios, de formación profesional y cursos de formación en empresas.
- A la vista de los resultados obtenidos, podemos afirmar que el perfilado automático permite obtener perfiles relevantes de la oferta curricular. Si bien podría mejorarse con algunas modificaciones en el diseño algorítmico, incluyendo alguna supervisión en el proceso de entrenamiento de modelos o, incluso, con una fase de postprocesado manual.

3.3.1 Análisis de títulos universitarios TICC

- Para el análisis de los títulos universitarios se hará uso de los planes de estudio disponibles en el Registro de Universidades, Centros y Títulos (RUCT).
- A 30 de septiembre de 2015 se detectaron 585 planes de estudios y 864 titulaciones universitarios relacionados con las TICC. Los grados repartidos entre Grados y Masters.

- En los planes de estudio, el 64,9% son de ramas de Ingeniería y Arquitectura, el 20% a la de Ciencias Sociales y Jurídicas, 9,9% a las ramas de Arte y Humanidades, y el 5% a ramas de Ciencias.
- Para las titulaciones, el 62,7% son de ramas de Ingeniería y Arquitectura, el 18,5% a la de Ciencias Sociales y Jurídicas, el 11,8% a las ramas de Arte y Humanidades, y el 6,9% a ramas de Ciencias.
- Los principales perfiles extraídos de la colección de planes de estudios universitarios se pueden agrupar en las siguientes 6 categorías:
    1. Informática, representando el 18% del corpus. Destacan dos perfiles, el primero caracterizado por los términos computadores, computación, informáticos, sistemas operativos, inteligencia, inteligentes artificial, diseñar, distribuidos, estadística. El segundo por los términos computación, computadores, informáticos, diseñar, algoritmos, empotrados, prestaciones, sistemas operativos, implementación, distribuidas.
    2. Comunicaciones, aglutina el 11% de las titulaciones del corpus. Los términos telecomunicaciones, electrónico, circuitos, señales, electrónica, telemática, transmisión, dispositivos, procesado y móviles, caracterizan al principal perfil de esta categoría.
    3. Contenidos o publicidad, representa el 33,5% del corpus, en la que se encuentran 6 perfiles, el más importante de los cuales se caracteriza por los términos audiovisuales, periodismo, producción, historia publicidad, televisión, radio, imagen, sem, informativos
    4. Electrónica, con el 12% de las titulaciones.
    5. Matemáticas, representa el 11% de las ofertas del corpus
    6. Otros, aglutinando a otros perfiles más genéricos y transversales.

3.3.2 Análisis de cualificaciones profesionales

- Para el análisis de cualificaciones profesionales se ha tomado como fuente de datos la información que publica en su web el Instituto Nacional de las Cualificaciones[1] (INCUAL).
- A 15 de julio de 2016 se han descargado 72 cualificaciones profesionales, el 43% relacionadas con las ramas de Artes Gráficos, el 32% de Informática y Comunicaciones, y el 25% restante de Imagen y Sonido.
- Los perfiles extraídos de las cualificaciones profesionales se pueden agrupar en tres categorías:
    1. Artes gráfica, los 8 perfiles que pertenecen a esta categoría totalizan el 54% de las titulaciones del corpus. El perfil más importante tiene un peso del 17%, y se caracteriza por los siguientes términos: tintas, encuadernación, primas, impresoras, tapas, cilindros, pre impresión, tirada, defectos y muestras.
    2. Imagen y sonido, representando el 19% del corpus, y compuesto por 4 perfiles. El perfil más importante representa el 7,8% del corpus, y está caracterizado por los términos: audiovisuales, cámaras, sonido, grabación, video postproducción, televisión, rodaje, movimiento y audio.

[1] INCUAL, http://www.educacion.gob.es/iceextranet/

3. Informática y Comunicaciones, que representa el 27%. Lo conforman 3 perfiles, el más importante de ellos, con un peso del 8,8%, se caracteriza por los términos: servidores, web, gestores, páginas, almacén, componente, multimedia, sistemas operativos, programación y mensajería

3.3.3 Análisis de cursos de formación en empresas

- Los perfiles extraídos de las ofertas de cursos formativos de empresas generan los siguientes categorías:
  1. Dos perfiles se refieren al uso de hojas de cálculo o Excel, totalizando el 22,9% de los cursos.
  2. El 19% de los cursos se agrupan en tres perfiles que están relacionados con las redes sociales y contenidos web.
  3. Perfiles relacionados con bases de datos hay 2, que totalizan el 13,8% de los cursos.
  4. Sobre procesamiento de texto e imágenes se detectan 3 perfiles, totalizando el 18% de los cursos.
  5. Un perfil se relaciona con la gestión de proyecto, aglutinando el 9,4% de los cursos.
  6. De los sistemas operativos se encuentra un perfil que totaliza el 6,6% de los cursos.
  7. El resto de perfiles se refieren a habilidades o características transversales a todos los cursos.

3.4 Análisis comparativo de oferta y demanda de profesionales TICC

- Se analiza el grado de ajuste entre la oferta de empleo y la oferta curricular en el sector TICC en España. Para ello, se realiza un análisis comparativo entre los perfiles extraídos para portales de empleo y oferta curricular, estimándose el grado de alineamiento entre unos y otros. Asimismo, se analiza qué perfil de oferta curricular encaja mejor con cada una de las ofertas de empleo, y viceversa.
- El objetivo final es hacer uso de los datos disponibles y de los perfiles extraídos para valorar hasta qué punto la oferta formativa en España responde a las necesidades reales del sector TICC, y qué aspectos formativos sería necesario reforzar por estar escasamente cubiertos en la actualidad.
- Se determina el grado de ajuste del contenido de los planes de estudios universitarios con las necesidades de las ofertas de empleo.

3.4.1 Ajuste de la oferta universitaria

- Se determina el grado de ajuste del contenido de los planes de estudios con las necesidades de las ofertas de empleo.
- Se puede ver los grupos de titulaciones que tiene mayor o menor demanda profesional.
  - Los perfiles formativos asociados con informática están demandados por casi todos los grupos de ofertas de empleo.
  - Hay perfiles formativos asociados con audiovisuales y periodismo, que solo son demandados por ofertas de marketing y comunicación.

- Otros planes formativos relacionados con las comunicaciones, procesado de señales y datos también tienen una alta demanda y se encuentran relacionados con diferentes perfiles de las ofertas de empleo.

3.4.2 Ajuste de la oferta de formación profesional

- Los perfiles que podríamos asociar, a la vista de los términos que los caracterizan, con Informática y Comunicaciones son los más demandados por las ofertas de empleo. Ambos están relacionados prácticamente con todos los perfiles de las ofertas.
- Perfiles menos demandados como los asociados con la rama de artes gráficas. Estos perfiles, a pesar de ser los que mayor peso adquirían en el corpus de formación profesional, son los que menos relación tienen con las ofertas de empleo, tanto en número de perfiles cubiertos como en grado de ajuste.

3.4.3 Ajuste de la oferta de cursos de formación en empresas

- Prácticamente todos los grupos de cursos que ha proporcionado el perfilado son demandados por un elevado número de los perfiles de las ofertas de empleo. Lo que indica que estos cursos están más relacionados con la demanda actual del mercado laboral.

3.5 Rankings de cobertura de los perfiles de ofertas de empleo

- Se ha medido el nivel de ajuste global entre demanda de profesionales (ofertas de empleo) y oferta de titulaciones, y generar una lista ordenada de perfiles según su cobertura.
- Respecto a la cobertura de perfiles de ofertas de empleo por parte de titulaciones universitarias, la oferta de empleo que mejor se ajusta a las titulaciones universitarias es la caracterizada por los términos: Informática, ingeniería, programación, inglés, francés, telecomunicaciones, ingeniero, equipo, sistemas e informático. Sin embargo esta oferta representa al 4,4% de las ofertas, ocupando la posición 12 de 20 dentro de las ofertas de Infojobs.
- Por el contrario, el perfil que aglutina a mayor número de ofertas de empleo ocupa la posición 20 de 20 en el ranking de ofertas con mayor ajuste con las titulaciones. Este perfil está caracterizado por los términos java, programador, j2ee, analista, spring, hibernate, javascript, cliente, aplicaciones y web_services, aglutinando el 8,52% de las ofertas.
- En el caso de la formación profesional, la oferta de empleo que mejor se ajusta a las titulaciones es la que se caracteriza por los términos seguridad, test, gestión, software, qa, análisis, ejecución, técnico, consultor y tester. Esta oferta aglutina al 2,64% de las ofertas de infojobs.
- La descripción de los cursos formativos en empresas tiene un formato mucho más alineado y con las ofertas de empleo, y cubre un temario mucho más específico, lo que explica una mayor frecuencia de terminología

tecnológica, y en última instancia permite medir de manera más fiable el ajuste entre dicha oferta formativa y las demandas del mercado de trabajo.