Ibermática

Extracción automática de conocimiento sobre OpenData (OpenMining / BigAnalytics)

Open Data es un nuevo paradigma en el que la tecnología actúa como facilitador en la publicación del conocimiento sito en la Administración Pública y los gobiernos. Así, se está configurando como una herramienta que permite realizar procesos de extracción, tratamiento y carga con múltiples fuentes de datos públicas, para generar informes o cuadros de mandos con un objetivo determinado.

Introducción

Open Data es un nuevo paradigma en el que la tecnología actúa como facilitador en la publicación del conocimiento disponible por la Administración Pública y los gobiernos. Así, se está configurando como una herramienta que permite realizar procesos de extracción, tratamiento y carga con múltiples fuentes de datos públicas, para generar informes o cuadros de mandos con un objetivo determinado. Así, podemos relacionar datos demográficos, políticos, censales, económicos, médicos, culturales, y pedir al sistema, que, en base a unos objetivos determinados (por ejemplo, el nivel de inversiones, indicadores de salud o resultados electorales), nos enseñe por qué se están produciendo dichos comportamientos y la propensión de que los mismos patrones se reproduzcan en el futuro.

Las contradicciones del "Open Data": "Open Government"

La tecnología, en estos tiempos modernos, usualmente va por delante de su aplicabilidad en distintos contextos funcionales, principalmente en la Administración Pública. El presente artículo no trata de ahondar en el estado del arte tecnológico, sino de demostrar cómo ciertas tecnologías, muy maduras en otros ámbitos (sanitario, bancario, marketing), pueden ser aplicadas al nuevo paradigma de Open Data, o publicación de datos abiertos por Administraciones Públicas, con el objetivo de mejorar la comprensión por parte de los "consumidores" de dicha información de forma directa, sencilla y rápida.

Tal y como describe Javier de la Cueva en su artículo "*Redefiniendo la isegoría: Open Data ciudadanos*", Open Data permite que un ciudadano (o administración) publique a coste cero información para que otros ciudadanos, pertenecientes a otro lugar del globo, puedan leerla a coste cero.

Sin embargo, para que esta posibilidad sea real, se deben dar las siguientes premisas:

- La liberación de los datos implica el sometimiento de los datos públicos a estándares abiertos obligatorios.
- Los datos dispuestos por los gobiernos deben ser lo más completos posibles, publicando toda la información en bruto, con la excepción de los datos relativos a la privacidad.
- Los datos puestos a disposición pública por los gobiernos deben ser fuentes primarias, y con un sentido de oportunidad, es decir, tan rápidamente como son reunidos y recogidos.
- Los datos deben de ser válidos, en un sentido estadístico, es decir, los valores intrínsecos a los mismos deben ser evidencias de la realidad en un porcentaje elevado (por ejemplo, se estima que los datos deben tener una confianza de un 80% en datos médicos).
- Calidad, veracidad e inmediatez.

La realidad es que, quitando contadas ocasiones, ninguna de las anteriores premisas se cumple. Por un lado, existe un estándar internacional para la publicación de los datos basado en tecnología semántica denominada "LinkedData", que tiene cinco niveles de "excelencia" y que no cumple casi ninguna administración. Por otro lado, los datos "expuestos" no son completos y es necesario "ahondar" en información adicional, principalmente de otras fuentes, o incluso manipulándola manualmente, para dar sentido a los datos publicados. Y finalmente, el coste que suponen los procesos de ETL (Extracción, Transformación y Carga) para las administraciones implica una clara pérdida de la oportunidad.

Quizás, el mayor problema en el coste de cumplimiento de estos requisitos es más profundo y resida en que no existen objetivos claros sobre qué conjuntos de datos publicar y principalmente para qué publicarlos. Se debe realizar un análisis previo e incluso una consulta a los agentes, principales "consumidores" de dicha información, con el objetivo de cuantificar claramente los "data sets" relevantes, priorizar y programar en el tiempo su publicación por orden de importancia,

gestionar correctamente la publicidad de dichas publicaciones, medir de forma objetiva y cuantitativa el retorno de la inversión (no siempre en parámetros económicos, sino de reputación e interés social) e incluso formar a la ciudadanía en las posibilidades de extracción y tratamiento de toda esta información. Sin tener los objetivos claros y los beneficios cuantificados, a pesar de estar muy avanzada tecnológicamente, la filosofía Open Data no tiene futuro.

Así, se está generando un fenómeno, por el cual los que realmente están liberando datos en formato "Open Data" son los ciudadanos mediante técnicas de transformación, extractando el conocimiento de forma normalizada, que el Estado ha sido incapaz de realizar en base a los criterios prefijados. Y en parte, la razón de este fenómeno es la contradicción que existe entre la "recomendación" de publicar datos en bruto (Estado), con la necesidad de extraer de dichos datos sólo aquella información que es relevante para las necesidades de una consulta concreta (Ciudadanos).

Y de esta forma, nace el concepto de Open Government. Open Government se refiere al uso de tecnología para fomentar la transparencia, participación y colaboración con los grupos de interés de la Administración Pública y los gobiernos. Aunque la expresión tiene raíces en el Siglo de las Luces y posteriormente en una declaración de Lincoln en 1863, Obama le dio un nuevo impulso a través de su "*Memorandum on Transparency and Open Government*" (2009). Para Obama, Open Government debe conducir hacia una Administración Pública y un gobierno más eficientes y efectivos. De alguna manera se trata de reinventar la gestión pública y el gobierno.

Conceptualmente, Open Government es la suma de Government 2.0 y Open Data. Es un nuevo paradigma en el que la tecnología actúa como facilitador de una transformación en la manera de gestionar la Administración Pública y los gobiernos, a través del fomento de la transparencia, la participación y la colaboración con los grupos de interés. Es decir, es la suma de la publicación de datos en bruto, en donde los grupos de interés (ciudadanos, empresas, gestores) son capaces de transformar dicha información en conocimiento publicable, recabando un beneficio final, bien renumerado, bien en conceptos de prestigio social. Como consecuencia de dicha aplicación, se obtienen resultados finales en términos de productividad, innovación y reputación/fidelización de dichos grupos de interés (ciudadanía).

Pero para que el "cliente final" pueda recibir en una aplicación "ad hoc" de forma directa las conclusiones de dicho conocimiento, se deben realizar una serie de pasos sobre la información base, como son:

- La normalización y diferenciación entre lo que son datos constantes, conceptos (denominados "URI"s en semántica) y lo que son los valores asociados a dichos conceptos, y que serán los que se puedan mostrar en base a gráficos o informes. Por ejemplo, Barcelona tiene una población de 1.621.537 habitantes se transforma en "dbpedia.org/page/Barcelona□dbpedia-owl:populationTotal□1621537 (xsd:integer)"

- Los datos deben estar relacionados, de forma que, a partir de ciertos "pasos" precalculados, podamos buscar relaciones entre distintos conceptos y sus valores. Por lo tanto, podremos relacionar el impacto, por ejemplo, de ciertas subvenciones sobre la renta de una determinada región, o sobre su efecto en enfermedades determinadas, si se solicita. Esto implica que hay que "normalizar" la información, según lo expresado en el punto anterior, pero además hay que "enlazar" dichos conceptos.

- Los datos deben ser relevantes, es decir, hay que "luchar" contra la cacofonía del ruido del entorno y presentar solamente aquella información que tiene relación con la "pregunta" que se quiere responder. Por ejemplo, si estamos buscando si existe relación entre las sentencias judiciales y su efecto sobre la violencia de género, "intuimos" que la información referente a los datos demográficos puede ser de interés, pero no así los datos meteorológicos, aunque igual nos sorprendíamos.

Como se puede sospechar, el "navegar" por la ingente cantidad de datos en bruto, para seleccionar los conjuntos de datos que "a priori" puedan tener relevancia con nuestro objetivo, enlazarlos, filtrarlos, normalizarlos, certificar su validez y presentarlos en un formato "usable", no es una labor trivial.

Y sin embargo, en el entorno médico, por ejemplo, existen millones de "filas" de información ya extractada en este formato, disponible de forma directa y accesible en función de las distintas preguntas que deseemos hacer. ¿Cómo es posible? Gracias a tecnologías de extracción, análisis y asociación de relaciones basadas en técnicas de Inteligencia Artificial, conjuntadas en lo que hoy en día se ha dado por denominar "Big Data" o "Big Analytics". El reto está en proporcionar dichas técnicas a los agentes del "Open Governmet", para que, de una forma desasistida, los datos en bruto se puedan transformar en conocimiento elaborado con el mínimo esfuerzo posible.

Big Data / Big Analytics

La palabra de moda: "Big Data"

Durante los últimos años, en el sector TIC, se ha pasado de la obsesión por "la nube" al foco en el "Big Data". No obstante, el término "Big Data" es relativo. Se emplea (según definición de Gartner) cuando los problemas de gestión y procesamiento de la información "superan en una o varias dimensiones la capacidad de las tecnologías tradicionales de gestión de información para respaldar el uso de este activo". Es decir, que los datos sólo son "Big Data" cuando no es posible gestionarlos o analizarlos.

Durante décadas, las TICs han salvado limitaciones conocidas, alojando datos en estructuras definidas o arquitecturas de almacenamiento. Con métodos que se basan en el indexado y los lenguajes primitivos, las bases de datos no tardan en volverse demasiado grandes para ser gestionadas. Pero, ¿qué pasaría si pudiésemos poner una sola matriz de memoria con una fila por resolución judicial?, o ¿crear mil millones de filas, una por cada sentencia, sus datos asociados (niveles de renta de los demandantes/demandados, perfil de los jueces, situación geográfica, información censal) y su conclusión? Sería ideal poder obtener respuestas a cualquier pregunta en segundos a través de una sencilla interfaz gráfica o simplemente a través de una Web accesible. Esta posibilidad es lo que denominamos "Big Analytics".

La tecnología de "Big Analytics", evolución de la ya clásica "Minería de Datos", se basa principalmente en la capacidad que tienen las máquinas de analizar correlaciones, relaciones, segmentaciones y procesos estadísticos en tiempo máquina ("sin descanso"), sobre un volumen de información ingente, tanto estructurada como no estructurada. Hay que tener en cuenta que el 80% de la información actual está en formato "textual" y hacen falta procesos de transformación de dicho lenguaje a un formato "normalizado". Por lo tanto,

"Big Analytics" aúna técnicas de procesado estadístico con técnicas de procesado de lenguaje natural, que además "encajan" a la perfección con la salida deseada que hemos comentado en el punto anterior, un formato semántico estructurado según al normativa de LinkedData.

"Big Analytics" se define como el proceso de descubrir los patrones de información interesante y potencialmente útil, inmersos en grandes fuentes de información dispersas con la que se interactúa constantemente. Internamente, es una combinación de procesos como:

- Extracción de datos.
- Limpieza de datos.
- Selección de características principales.
- Algoritmos de clasificación y predicción.
- Análisis de resultados.

Estas plataformas exploran una gran cantidad de datos y, mediante su análisis, explican qué indicadores tienen correlación con ciertos objetivos o preguntas realizadas, y además cuáles son las reglas que modelan dichos comportamientos. Una vez extraídas dichas reglas, es posible predecir posibles tendencias o comportamientos futuros dentro de una entidad, permitiendo al usuario final "comprender" la lógica de lo que los datos "dicen", y los datos "nunca mienten", y en base a ello, poder tomar decisiones, en unos casos, o poder publicar noticias basadas en la investigación de los datos (como es el caso del "periodismo de datos").

La diferencia de estas técnicas con las clásicas estadísticas reside, principalmente, en que las técnicas estadísticas se centran en técnicas confirmatorias y "Big Analytics" en técnicas de descubrimiento. Así, cuando el problema al que pretendemos dar respuesta es refutar o confirmar una hipótesis, podremos utilizar ambas ciencias. Sin embargo, cuando el objetivo es meramente exploratorio (para concretar un problema o definir cuáles son las variables más interesantes en un sistema de información) aumenta la necesidad de delegar parte del conocimiento analítico a técnicas de aprendizaje. Así, "Big Analytics" se utilizará cuando no partimos de supuestos de inicio y pretendemos buscar algún conocimiento nuevo y susceptible de proporcionar información novedosa en la toma de decisiones.

En el caso de datos públicos, y siguiendo con la premisa de publicación en bruto por parte de la Administración, existe una alta dimensionalidad del problema. Cuantas más variables entren en el problema, más difícil resulta encontrar una hipótesis de partida interesante o, aun cuando se pudiera hacer, el tiempo necesario no justificaría la inversión. En ese caso, utilizar técnicas de minería de datos como árboles de decisión nos permitirá encontrar relaciones inéditas para luego concretar la investigación sobre las variables más interesantes y, al contrario que en la estadística, cuantos más datos tengamos, mejor solucionaremos el problema.

No es el objeto de este artículo ahondar en la algoritmia interna de estas plataformas, pero como un punto general, se puede decir que se trabaja en distintas fases:

- **Clustering**: Es un planteamiento que intenta identificar las características distintivas entre los conjuntos de registros y el lugar en grupos o segmentos. Este proceso es a menudo la intensificación de punto de partida para la minería de datos, ya que conduce a la exploración de relación. Este proceso en particular es un candidato obvio para la segmentación de clientes por agrupación de similitudes.
- **Asociación**: Aquí se encuentran las reglas que le permiten correlacionar la presencia de un conjunto de elementos con otro conjunto. Este método ha demostrado ser eficaz en el comercio minorista, donde el análisis de la cesta de la compra le ayuda a encontrar que algunos artículos son siempre comprados en el mismo tiempo. Si usted puede encontrar los patrones de compra natural de un cliente, puede utilizar ese modelo para ayudar a comercializar su producto. El resultado de esta asociación es una lista de afinidad de productos.

- **Asociación secuencial**: Patrones relacionados con el tiempo. Este método busca los vínculos que relacionan estas pautas secuenciales. La idea es utilizar los datos asociativos como una cuenta de cheques o de un acontecimiento vital para unir una secuencia de acontecimientos en una serie de tiempo. La vida activa que precede a sus compras y las compras de precursores se encuentran a menudo con esta metodología.

La curva de aprendizaje

Una vez demostradas las capacidades de esta tecnología sobre el tratamiento masivo de datos, el problema es cómo "transferir" dichas capacidades a los verdaderos agentes de la transformación y reutilización de datos abiertos, es decir, a los ciudadanos o empresas que los modelizan. Sin embargo, el problema no es tal, ya que existen múltiples plataformas analíticas avanzadas, muchas de ellas de software libre disponibles por la "comunidad" para su uso.

Estas plataformas tienen distinto grado de usabilidad y accesibilidad, pero las hay que, desde entornos Web, son capaces de permitir a los usuarios la subida de datos y la bajada de las reglas que explican dichos datos, en base a reglas que aclaran de forma visual por qué un indicador está ocurriendo en determinados casos y en otros no. Otro tipo de plataformas (RapidMiner, Knime), tienen una curva de aprendizaje un poco más escarpada, pero es perfectamente viable, con pocas jornadas de entrenamiento, que una persona no informática sea capaz de generar sus propios modelos para obtener las reglas que modelan sus datos.

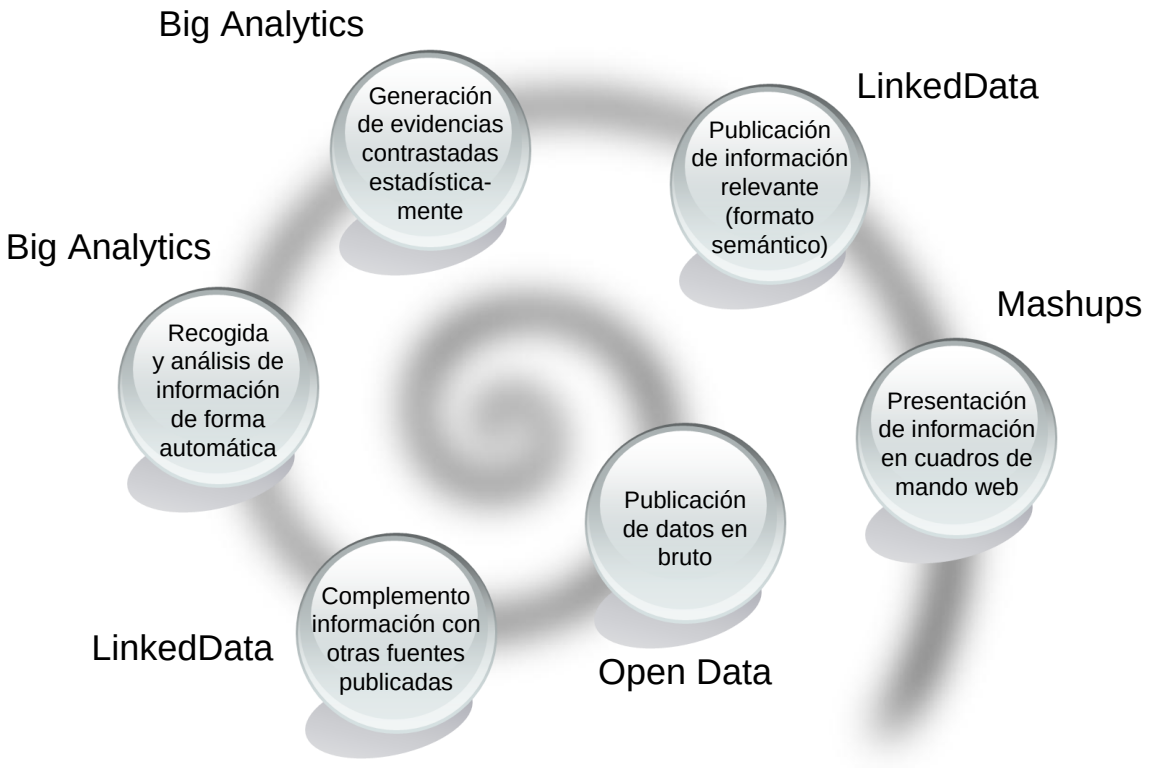
Por lo tanto, una vez más, es evidente que la tecnología supera al "negocio" en cuanto a facilidad de uso, pero muchas veces dicha funcionalidad no es divulgada, en primer lugar, por intereses particulares, en los que determinados nichos de negocio (empresas de marketing, consultorías en analítica avanzada) pretenden "mantener" el conocimiento de estas tecnologías en nichos cerrados para poder seguir operando como "gurús tecnológicos" especializados en "siglas de "tres letras", aparentemente inaccesibles para el resto de la sociedad, cuando las plataformas reseñadas anteriormente tienen una filosofía totalmente abierta y accesible. De hecho, la mayoría de las herramientas disponibles utilizan el mismo tipo de algoritmos, la misma base de métodos estadísticos o son variaciones "sutiles" de los métodos generales.

Big Analytics vs Open Data

Según Gartner, "Big Data nos hace más listos, pero Open Data nos hará más ricos". Efectivamente, tal y como hemos reseñado, Big Data nos permite comprender qué hay en nuestros datos, justo aquello que es relevante para nuestras consultas y sólo aquello relevante, así como el porqué está ocurriendo, mientras que Open Data permite a los agentes "publicar" dicha información transformada en conocimiento, para su propio beneficio.

La conjunción de ambas tecnologías permite minimizar el coste de análisis, transformación y extracción de la relevancia, aumentando, por lo tanto, el beneficio de la "venta" de dicho conocimiento a terceros. De esta forma, las Administraciones "delegan" su esfuerzo de desarrollo a agentes externos, siendo su única responsabilidad la de publicar datos en bruto.

El flujo del proceso es simple: Publicación de datos en bruto por parte de la Administración (Open Data), complemento de dicha información con otras fuentes ya publicadas (LinkedData), recogida y análisis de información de forma automática (Big Analytics), generación de evidencias contrastadas estadísticamente (Big Analytics), y publicación de la información relevante en formato semántico (LinkedData), más la presentación de dicha información en cuadros de mando Web (Mashups). Todo ello conformaría el ciclo en espiral de la gestión del conocimiento en un entorno de "Open Government".



Casos de uso

Datos abiertos y retos a resolver

Actualmente, en el ámbito del Gobierno y Justicia, a nivel internacional se están publicando datos que permiten a terceros, como ya se ha explicado anteriormente, realizar un tratamiento de los mismos, "mashups", para ofertar conocimiento a problemáticas concretas. Por reseñar algunas, ya que no es el objetivo de este artículo, se puede nombrar a la Administración Británica, por ejemplo, muy activa en sus publicaciones, permitiendo la creación de aplicaciones del tipo "nivel de criminalidad por condados", o tipologías de crímenes "a tu alrededor", o reseñas legales relacionadas con alimentos y leyes.

Así mismo, el portal legislation.gov.uk impulsado por los Archivos Nacionales, órgano dependiente del Ministerio de Justicia del Reino Unido, es uno de los proyectos pioneros en la utilización de los criterios de normalización mencionados anteriormente para abrir los datos públicos. Este portal ofrece toda la información sobre la legislación británica a través de un sitio web con servicios adecuados para la búsqueda de información por parte de los ciudadanos y, a su vez, publica toda la información de forma estructurada y marcada de forma semántica para que pueda ser procesada automáticamente.
Por otro lado, Estados Unidos, líder en la publicación de datos en abierto, tiene su propio portal de justicia con multitud de sets de datos relacionados en su web www.justice.gov/open/data.html.

A nivel nacional, con respecto a datos judiciales, el Consejo General del Poder Judicial publica documentación sobre sentencias y estadísticas dentro de su portal poderjudicial.es, con un trabajo muy elaborado de agregación, filtrado y un amplio conjunto de indicadores. En la bibliografía se encontrarán ejemplos de portales de datos públicos en distintas comunidades.

Pero lo realmente importante es analizar si dichos datos resuelven la problemática de deficiencia de información ciudadana. Y éste es un ejercicio que están realizando asociaciones como la Red Temática Española de LinkedData que, con datos Open Data de distintas administraciones, intentan publicar datos estructurados en abierto para dar respuestas a cuestiones como la posible relación entre empleo y desahucios, la calidad del aire en una zona geográfica y su relación con la atención sanitaria, la calidad del aire y su relación con la mortalidad, etc.

Según se vayan teniendo datos "ad hoc" al entorno judicial, con los datos que se tienen a nivel estadístico (INE), autonómico y nacional, y las estadísticas judiciales, otras cuestiones a realizar en el futuro de interés para la comunidad son, por ejemplo, el análisis del sentido de los cambios legislativos y su incidencia en la litigiosidad para determinar costes para la administración (justicia gratuita, necesidades de personal, participación de peritos judiciales…), la relación entre los costes de los peritajes judiciales y los perfiles de las temáticas, el cálculo predictivo de los costes procesales por tipología, el ámbito geográfico y el perfil de los intervinientes, el análisis de la Justicia criminal (transfronteriza), los tiempos de resolución y las razones de los distintos retardos, etc.

Pero todas estas preguntas quedarán sin respuesta si no aunamos los dos mundos, el del análisis automático sobre los datos en bruto y la publicación de los indicadores relevantes y sus "reglas" de funcionamiento, en un entorno abierto y accesible.

Un ejemplo: ¿Es posible conocer la intención de voto en base a datos abiertos?

Con los datos censales de unas elecciones en particular ¿qué podemos obtener? Básicamente, conocer el porqué de los resultados electorales y no quedarnos simplemente en lo que ha ocurrido (pasado estadístico), sino intentar comprender las razones generales de dichos resultados y poder aplicar dicho descubrimiento al futuro.

Comenzamos tomando los resultados electorales de una provincia, por ejemplo, Bizkaia, y para ello, recogemos los datos de los resultados de las Elecciones Municipales de 2007, extraídos de los datos publicados por OpenDataEuskadi. En esta relación, tenemos datos por mesa censal del número de votos por partidos, nulos y blancos. Estos son los datos clásicos de los resultados electorales y con ellos podemos presentar cientos de estadísticas, a nivel de Municipio, Distrito y Sección Censal, incluso a nivel de mesa electoral, del número de votantes, los votos nulos, los válidos, los partidos ganadores por dichas secciones, etc. En definitiva, los informes estáticos de los que hablamos en la introducción.

Ahora bien, esta información es meramente descriptiva, así que la pregunta que debemos hacernos ahora es ¿existirá algún patrón de comportamiento que explique la intencionalidad del voto, en base a algún indicador adicional? Pues bien, como primer paso, necesitamos agregar información pública de otros lugares, referente a los datos demográficos, económicos y otros.

Intuitivamente, los datos demográficos asociados a las Secciones Censales pueden explicar comportamientos relativos a los movimientos de masas sociales en los municipios estudiados. Es decir, una Sección Censal agrupa grupos poblacionales, en teoría, más o menos homogéneos a nivel cultural, educacional y económico. Además, por Sección Censal y gracias al Instituto Nacional de Estadística, tenemos datos sobre los valores de población por sexo en intervalos de edad en dicha Sección, además de los totales de personas nacidas en la misma Comunidad Autónoma o distinta, incluso los que siendo de la misma Comunidad Autónoma son de la misma provincia o diferente, incluso nacidos en el mismo municipio o distinto, dentro de la misma Provincia, separados por sexo.

En cuanto a los datos relacionados con los Municipios, tenemos información sobre el número del desempleo registrado, tanto por hombres como por mujeres, así como los datos acumulados de sus poblaciones, diferenciadas por sexo.

Toda esta información la hemos recogido de la página del INE (Revisión del Padrón Municipal 2007, Datos por municipios) y la hemos incorporado a nuestra base de datos con relación a la Sección Censal y al Municipio. Éste es un ejemplo de agregación de distintas fuentes de información pública. Al final, tenemos una única tabla enlazada de 126 indicadores que incluyen toda la información acumulada por Sección Censal y Municipio.

Una vez enlazados todos estos datos, sólo queda un paso más: introducirlos en la máquina de "Big Analytics", definir un objetivo y ejecutar el proceso de modelado. El objetivo está claro, queremos conocer qué pautas se siguen, si es que existen, para determinar cuál es el partido más votado, es decir, el objetivo sería el campo "Partido Ganador", independientemente de cuál sea en cada Sección Censal.

Es importante anotar que no estamos realizando un estudio concreto de un Municipio (como podría ser, por ejemplo, el estudio concreto de Bilbao), sino de toda la provincia de Bizkaia. Estamos buscando reglas generales de comportamiento, que afecten a todos los indicadores, sin realizar un filtro previo por Municipio. Si hay correlaciones y no estamos mezclando peras con manzanas, el sistema las extraerá, si no, no será capaz de concluir ninguna regla o ningún modelo con una confianza lo suficientemente segura como para dar por válido el modelo.

Lo primero que hacemos es ejecutar un proceso de correlación que nos indique qué campos son los más relevantes con respecto a nuestro indicador objetivo, es decir, el "Partido Ganador".

Como resultado, lo que más relevancia tiene a la hora de determinar el partido ganador no es la Sección Censal, sino el Municipio (NOMBRE), seguido directamente, pero con muy poca correlación, por la relevancia de la población nacida en Gipuzkoa y Álava y que vota en Bizkaia, y por los datos del paro.

En cuanto a la explicación de dicha relevancia, se obtienen una serie de reglas del tipo:

Si Paro registrado, Mujeres > 423
y NOMBRE = Amorebieta-Etxano
y además MujNacDistComAut (Mujeres Nacidas en Distinta Comunidad Autónoma) > 179,5
entonces –> PSE_EE_PSOE

Es decir, el paro registrado en las mujeres (a nivel municipal) es un indicador clave en el comportamiento de la intención del voto, seguido de la procedencia del nacimiento de las mismas mujeres que votan en una determinada mesa censal.

Si analizamos el caso de Bermeo, las reglas son diferentes, puesto que el sistema es lo suficientemente flexible para determinar patrones diferentes reflejando comportamiento diferente:

Si NOMBRE = Bermeo
y Muj_25-29 > 45, entonces –> EAJ_PNV
pero si Muj_25-29 <= 45 entonces –> EA

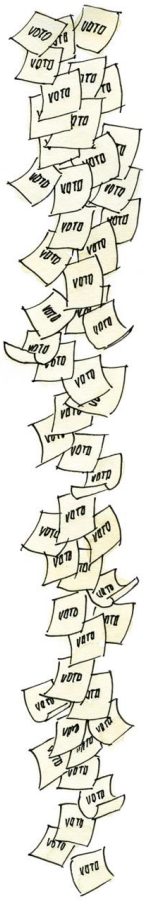
En este caso, la diferencia de comportamiento viene dada por la edad de las mujeres que votan, dato que también está incorporado a nivel de Sección Censal. Es decir, si en Bermeo, en una Mesa Electoral, hay más de 45 mujeres entre 25 y 29 años, entonces, en esa mesa saldrá elegido el partido EAJ_PNV, pero si hay menos de 45 mujeres en ese ratio de edad, ganará EA.

Por otro lado, si EA quiere mejorar sus resultados, debería enfocar su campaña en Bermeo a las mujeres entre 25 y 29 años, puesto que no parece que ese "perfil" poblacional en Bermeo sea de sus siglas, y si consigue captar ese foco de atención, teniendo en cuenta que el margen de diferencia en votos es muy pequeño, es posible que gane en las próximas elecciones en dichas mesas electorales.

Viéndolo en el mapa, incluso sabemos en qué zonas de Bermeo habría que realizar esta acción:

Como indicábamos al principio, estas reglas deben tener un componente de validez, que también nos lo da el sistema, en forma de un nivel de confianza porcentual de cada regla y del conjunto del modelo. Si miramos la confianza de este sistema en cuanto a predicción, vemos que el modelo es estable con una confianza de un 79,4%, es decir, que el sistema es capaz de acertar con las reglas expuestas anteriormente en un 79% de los casos y, por lo tanto, predecirlas con ese grado de fiabilidad.

Como conclusión, el análisis automático de la unión de datos demográficos con los resultados electorales, proporciona a los gestores de campañas una información adicional y enriquecida sobre las causas de dichos resultados, pudiendo a futuro predecir, comprender e intentar canalizar esfuerzos en aquellos segmentos de la población a los que realmente debe influir para mejorar sus resultados en próximas campañas.



Bibliografía

- Sir Tim Berners Lee en noviembre de 2010, durante una charla a "The Guardian", transmitió el mensaje de que "*Analizar los datos es el futuro de los periodistas*" http://www.guardian.co.uk/media/2010/nov/22/data-analysis-tim-berners-lee
- European Public Sector Information (PSI) Platform14 EPSI: http://www.epsiplatform.eu/
- Open Knowledge Foundation: http://okfn.org/
- Mapa mundial de Open Data: http://datos.fundacionctic.org/sandbox/catalog/faceted/
- Open Data Euskadi: http://opendata.euskadi.net/
- Datos Abiertos de Asturias: https://www.asturias.es/portal/site/Asturias/
- Dades Obertes: http://dadesobertes.gencat.cat/
- Datos Abiertos Zaragoza: http://datos.zaragoza.es/
- Datos Abiertos Gijón: http://datos.gijon.es/
- GeoLinkedData: http://geo.linkeddata.es/
- Case Study: Use of Semantic Web Technologies on the BBC Web Sites: http://www.w3.org/2001/sw/sweo/public/UseCases/BBC/
- http://apunteselectronicos.blogspot.com.es/2011/01/del-opendata-al-openservices.html, Foros de opinión y discusión abierta promovidos desde expertos de la AGE más dinámicos en España en temas de Open Data.
- Apertura y reutilización de datos públicos, Rubén Martín, CTIC (2011).
- World Wide Web Consortium, Web Service Semantics - WSDL-S, Rama Akkiraju, Joel Farrell, John Miller, Meenakshi Nagarajan, Marc-Thomas Schmidt, Amit Sheth, and Kunal Verma, Authors. 7 November 2005.
- Harms, Sherri K; Deogun, Jitender, S. Sequential Association Rule Mining with Time Lags. Journal of Intelligent Information Systems, 2004
- Papasalouros A., Kotis K., Kanaris K., (2008), Automatic generation of multiple-choice questions from domain ontologies. IADIS eLearning 2008, Amsterdam
- Budzik, J. & Hammond, K. (1999). Watson: anticipating and contextualizing information needs. Proc. ASIS, 727-740

Garantía Ibermática

Ibermática es una de las principales compañías de servicios en Tecnologías de la Información (TIC) del mercado español. Creada en 1973, su actividad se centra en las siguientes áreas: Consultoría TIC, servicios de infraestructuras, integración de sistemas de información, outsourcing e implantación de soluciones integradas de gestión empresarial. Asimismo, está presente en los principales sectores de actividad: finanzas, seguros, industria, servicios, telecomunicaciones, sanidad, utilities y administración pública, donde ofrece soluciones sectoriales específicas. Completa su oferta con soluciones tecnológicas como Business Intelligence, ERP y CRM, gestión de procesos (BPM), recursos humanos, movilidad, gestión de contenidos (ECM), Social Business / Gov 2.0, gestión de personas (HCM), Arquitecturas SOA, trazabilidad, accesibilidad, seguridad e inteligencia artificial, así como servicios Cloud Computing.

Con 40 años de actividad en el sector de las TIC, Ibermática se ha consolidado como una de las primeras empresas de servicios de TI de capital español. Actualmente agrupa a 3.278 profesionales y representa un volumen de negocio de 247,7 millones de euros.

Persona de contacto:

Aitor Moreno
Responsable de Inteligencia Artificial
ai.moreno@ibermatica.com

Pedro de la Peña
Investigador i3B
pm.delapena@ibermatica.com

Tel.: 902 413 500
www.ibermatica.com

Ibermática
Tecnología y Conocimiento