La Georreferenciación como Herramienta para la Gestión del Conocimiento en el Contexto de un Observatorio Regional de Desarrollo de la Ingeniería en Sistemas de Información/Informática

Fabiana María Riva, Martin Abbatemarco, Alejo Cervino, Nicolás Pereira
Departamento de Ingeniería en Sistemas de Información
Facultad Regional Rosario
Universidad Tecnológica Nacional
E. Zeballos 1341, 2000 Rosario, Argentina
fabianamriva@gmail.com, abbatemarco.martin@gmail.com,
alejocervino@hotmail.com , npereira@frro.utn.edu.ar

Abstract

La obtención de indicadores que permitan la toma de decisiones estratégicas por parte de los sectores que conforman el triángulo de Sábato - Universidad, Estado, Industria -, en torno al desarrollo y evolución de las Tecnologías de Información y Comunicaciones, Software y Servicios Informáticos (TIC-SSI), y su aporte a las cadenas productivas transversales, ha sido una de las metas propuestas para el planteo del Observatorio Regional de Desarrollo de la Ingeniería en Sistemas de Información/Informática (IISI.d.r.O). En este sentido, se ha definido la Infraestructura Tecnológica, incluyendo el desarrollo de productos para la captura y sistematización de datos, así como para el cálculo de indicadores que sustentarán la operación de IISI.d.r.O. Sin embargo, y para avanzar en la utilización efectiva de la información obtenida, es necesario convertir al sistema de información en un recurso estratégico esencial para la gestión del conocimiento. En particular en este trabajo, y tomando como relevante la información obtenida de los conocimientos y competencias que demanda el mercado laboral en torno a las TIC-SSI, se analiza el desarrollo de un prototipo preliminar para la automatización de la recuperación y tratamiento de la mencionada información y la aplicación de la georreferenciación como herramienta, no sólo de visualización sino como disparadora para el descubrimiento de nuevas perspectivas de análisis de la información para la Gestión del Conocimiento que, sin el uso de estas tecnologías, permanecerían invisibles.

Palabras clave: *Observatorio - Competencias - Georreferenciación - Gestión del Conocimiento*

1. Introducción

El Observatorio Regional de Desarrollo de la Ingeniería en Sistemas de Información/Informática (IISI.d.r.O) tiene como finalidad el diseño, construcción e implementación de una plataforma tecnológica integrada y abierta que recopile, analice y suministre información sustantiva en torno al desarrollo y evolución de las Tecnologías de Información y Comunicaciones, Software y Servicios Informáticos(TIC-SSI) y su aporte a las cadenas productivas transversales, para atender a las necesidades de los sectores que conforman el Triángulo de Sábato: Universidad - Estado - Industria[1].

Habiendo ya analizado la inexistencia de Observatorios en el sentido que pretende darse a IISI.d.r.O.[1] y definida la metodología para el trabajo en el Proyecto[2], basada en las metodologías ágiles y enmarcada en un ciclo de mejora continua, se refinaron los objetivos específicos que determinaron los productos a desarrollar, comenzando así a delinear el diseño de la Infraestructura Tecnológica para su operación.

Algunos de los objetivos específicos más relevantes tomados en cuenta desde el inicio del Proyecto han sido:

- Identificar perfiles y competencias demandadas por la Industria SSI que, en relación a los conocimientos y competencias derivadas de la formación académica, pueden ser utilizadas como indicador clave para derivar las necesidades de formación a cubrir por parte de la Universidad.
- Reconocer requerimientos de investigación pertinentes vinculados a la carrera, región donde se desarrollan y sectores transversales de aplicación de forma tal de mantener actualizadas las temáticas de investigación prioritarias.

- Individualizar cuestiones referidas a la movilidad laboral de los egresados o estudiantes de la carrera y su relación con las competencias de egreso y profesionales adquiridas.

- Conocer el mapa de distribución de desarrollos de productos y servicios para la industria específica y transversales, necesidades de infraestructura, recursos humanos, financieros y de competitividad en relación a otros mercados, que servirán al Estado para diseñar políticas de acción.

- Analizar la existencia de segmentos de mercados no explotados, necesidades de las cadenas productivas transversales y existencia de recursos humanos con las competencias necesarias para la implementación de tecnologías, que servirán a la industria para mejorar sus propuestas.

Los avances de IISI.d.r.O. orientaron la utilización del constructo **Competencias** como base que sirvió de vehículo para el planteo de una "Red para el Análisis comparado de competencias en la trama productiva de la Industria del Software y Servicios Informáticos (SSI)"[3]. Esta línea de trabajo tiene su origen en el estudio de las competencias genéricas y específicas demandadas por el mercado laboral a partir de la sistematización de búsquedas en medios gráficos[4] y de la realización de encuestas a Empresas[5] y a Estudiantes y Egresados que realizan su actividad en la cadena productiva SSI[6].

El planteo de la mencionada Red en función de los objetivos planteados supone, a grandes rasgos, la obtención de información relevante asociada a actividades de investigación, entendidas como aquellas que producen nuevos conocimientos y fomentan la fragmentación y expansión del mismo, actividades profesionales reservadas a los títulos y estándares de acreditación de las carreras de Ingeniería en Sistemas de Información e Informática[1], y conocimientos y competencias demandadas por el mercado laboral SSI. El sustento tecnológico de esta Red es una Base de datos de Grafos[7] con el objeto de determinar las relaciones entre los diferentes nodos (conocimientos y competencias) que la conforman y su nivel de influencia.

Particularmente, el presente trabajo surge de las diferentes problemáticas que se suscitaron de la recolección de los datos para la identificación de perfiles y competencias demandadas por la Industria SSI, requerida no solo para proveer de datos a la Red sino para cumplir con la meta de mantener la periodicidad de las observaciones, objetivo inherente a un observatorio. Entre las problemáticas caben destacar la escasa participación en las encuestas realizadas, la variación en las prácticas de reclutamiento de las empresas SSI, además de la merma en las publicaciones de las fuentes utilizadas.

En la búsqueda de nuevos mecanismos para cumplir con el objetivo, las fuentes seleccionadas esta vez fueron las redes sociales virtuales e Internet lo que derivó en el análisis de cuestiones técnicas, legales y éticas para su tratamiento [8, 9].

Las oportunidades de obtención de información brindadas por las redes sociales e Internet, en particular para este trabajo de las páginas de ofertas laborales, mejoraron tanto en volumen, variedad y extensión geográfica, orientando al equipo de IISI.d.r.O. al análisis de nuevas metodologías para la sistematización y tratamiento de los datos y a la utilización de nuevas herramientas para mejorar el análisis abriendo nuevas perspectivas y oportunidades de conocimiento, facilitando a su vez nuevas interpretaciones. En este sentido, la incorporación de la dimensión espacial favorece la extensión del estudio, inicialmente pensado para la Región Rosario, a todas las regiones donde se ubican las carreras de Ingeniería en Sistemas de Información/Informática, ayudando a la incorporación de estudios sobre las externalidades de conocimiento que específicamente se producen o deben producirse en cada una de ellas.

El presente trabajo plantea, entonces, el diseño de un prototipo preliminar que, de la sistematización y tratamiento automatizado de los datos obtenidos de Internet, en particular de las páginas de ofertas laborales para la identificación de perfiles, conocimientos y competencias demandados por la Industria SSI, y de la aplicación de georreferenciación al estudio, permita el descubrimiento de nuevas fortalezas y de debilidades a mitigar para cumplir con los objetivos de IISI.d.r.O.

A partir de los objetivos planteados para el trabajo y de la metodología de trabajo en el Proyecto se desarrollaron las siguientes fases:

- Análisis de Técnicas y Herramientas
- Desarrollo de funcionalidades
- Prueba del Prototipo

2. Análisis de Técnicas y Herramientas

2.1. Aplicación de la Georreferenciación al Estudio de la Trama Productiva SSI

En la búsqueda de metodologías que favorezcan la generación de conocimiento basado en las variables que se pretenden analizar en el contexto de IISI.d.r.O., surge, como

[1] Contenidos curriculares básicos, carga horaria mínima, criterios de intensidad de la formación práctica y estándares para la acreditación de la carrera de Ingeniería en Sistemas de Información. Ministerio de Educación. Resolución Ministerial Nro.786 del 26/05/2009 (publicada en el Boletín Oficial Nro. 31.667 del 4/06/2009). Disponible en: `http://www.coneau.edu.ar/archivos/Res786\_09.pdf`.

se ha mencionado en la Introducción, la oportunidad de incorporar la dimensión espacial, inicialmente como técnica empírica, que permita no sólo la visualización regional sino la determinación de nuevas variables y elementos a tener en cuenta.

Autores [10, 11, 12, 13] que trabajan la incorporación de la dimensión espacial a estudios sociales y económicos definen algunas cuestiones que se toman en consideración para el desarrollo del trabajo.

En principio se dirá que la metodología a aplicar alude a lo definido como análisis socioespacial con Sistemas de Información Geográfica(SIG) cuya base conceptual es la teoría de la geografía.

Los "SIG pueden ser entendidos como procedimientos técnicos y metodológicos que permiten por un lado tratar la espacialidad de los datos y por otro favorecer el estudio de la realidad desde enfoques multidimensionales e integrados, como son el tiempo, el espacio y las personas que interactúan con el territorio en un momento determinado"[14].

El principio de organización de los SIG posibilita la incorporación de diferentes capas que podrán ser analizadas en forma integrada identificando entidades y atributos temáticos. En función de esto el resultado será un **modelo vectorial** representado por puntos en el espacio de forma tal de modelar un fenómeno discreto.

Para el desarrollo del módulo de georreferenciación es necesaria la selección de una herramienta para la lectura, procesamiento y presentación final de la información obtenida en un mapa. Se decidió utilizar QGIS en su versión 2.18 [15]. QGIS es un sistema de información geográfica libre y de código abierto que permite la creación, edición, visualización, análisis y publicación de información geoespacial. La disponibilidad de documentación en línea, junto con su amigable interfaz de usuario y la fácil y rápida integración con Python, lenguaje de programación seleccionado para el desarrollo, fueron las principales razones que llevaron a la utilización de esta herramienta por sobre otras disponibles.

Si bien en este trabajo se focaliza sobre las ofertas laborales, donde las entidades serán los avisos y los atributos serán los perfiles, conocimientos y competencias demandados por la Industria, se han identificado además como relevantes al estudio de la trama productiva SSI las siguientes capas: empresas SSI para trabajar sobre las tecnologías aplicadas en el desarrollo de sus productos y sectores transversales que determinarán nuevas competencias, oferta académica universitaria, conocimientos y competencias que se derivan de actividades de investigación y oferta de profesionales y su movilidad.

2.2. Procesamiento de Ofertas Laborales

En este sentido se torna necesario resolver dos cuestiones fundamentales: la recolección de datos de las páginas de ofertas laborales y el procesamiento para la detección automática de perfiles, conocimientos y competencias relevantes como atributos para luego proceder a su georreferenciación.

Se avanzó en el estudio de métodos de extracción de datos de páginas web ya iniciado en trabajos anteriores[8] y con posibilidades de acceder a las páginas de ofertas laborales seleccionadas, cuyas APIs no proveen la totalidad de los datos necesarios para el trabajo. A partir de estudios exhaustivos de las técnicas existentes planteado por diferentes autores[16, 17, 18], se consideraron las particularidades de las páginas a analizar, detectando que su estructura era lo suficientemente sencilla para aplicar técnicas ad-hoc para la extracción de datos. Algunas problemáticas evaluadas a partir de esta decisión son, por un lado, la necesaria actualización requerida de las técnicas desarrolladas en función de la posibilidad de cambio de la estructura de las páginas y, por otro, las políticas de uso de las páginas de ofertas laborales, que limitó la posibilidad de selección de las mismas. Estas cuestiones serán abordadas con posterioridad en el Proyecto mediante la utilización de métodos automatizados o semiautomáticos para la modificación[16] de las técnicas y el planteo de convenios para un acceso más abarcativo y sustentable en el tiempo a las páginas seleccionadas.

Una vez obtenidos los datos iniciales fue necesario aplicar técnicas para el reconocimiento de perfiles, conocimientos y competencias existentes en el cuerpo de los avisos recuperados. Todas estas actividades tienen como objetivo final completar los datos del modelo base desarrollado en el Proyecto (Figura 1).

Las técnicas existentes para esta actividad abarcan desde técnicas de procesamiento de lenguaje natural y minería de texto hasta métodos de clasificación y etiquetado automáticos para llegar al descubrimiento del conocimiento al que alude la minería de datos, con métodos de aprendizaje supervisado y no supervisado[19].

Frente al trabajo a realizar y si bien se cuenta de trabajos anteriores[4] con listas de conocimientos y competencias asociados a perfiles, la aplicación de los últimos métodos mencionados requiere, no sólo de la estructuración de la fuente, sino de un período de entrenamiento y un conjunto de entrenamiento adecuados con el que aún no se cuenta a esta etapa del proyecto[20].

En cuanto a la aplicación de la minería de texto, el principal desafío que se enfrenta es la ambigüedad del lenguaje natural, cuestión que se torna más compleja ante la inexistencia de un corpus en el dominio específico que facilite la precisión y consistencia en el procesamiento.

Al analizar las fuentes seleccionadas pudo detectarse la inexistencia de complejidades semánticas, concepciones diferentes que puedan derivarse de una misma frase, ni complejidades sintácticas. Sólo se pueden encontrar posibles inconvenientes como la enumeración sucesivas de tecnologías

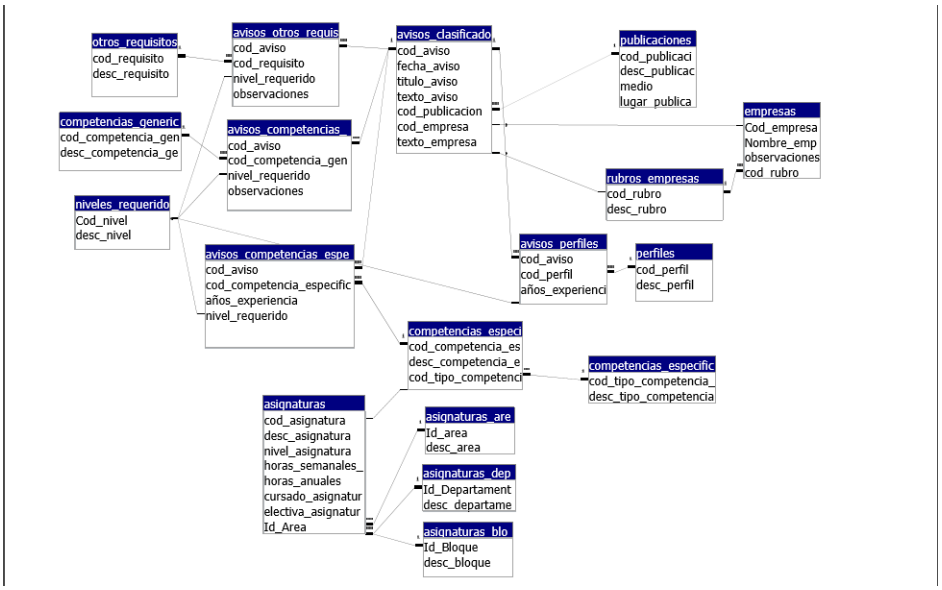

Figura 1. Modelo de Datos para la Sistematización de Demandas de Perfiles, Conocimientos y Competencias

en una misma frase, por ejemplo "con conocimientos en Ecmascript 5 o 6". Sin embargo, y considerando que generalmente las tecnologías se indican con versiones asociadas y su modificación es constante, no tendrá relevancia al estudio la individualización de las mismas.

Obtenidos los datos de las fuentes seleccionadas, se procede a un pre-procesamiento que refiere a la limpieza para eliminar etiquetas o caracteres de marcado que no aporten información útil para las siguientes fases, así como la eliminación de información redundante o errónea derivada del proceso de extracción.

Seguidamente, se selecciona el método de clasificación a utilizar que, para el presente trabajo, consistió en el uso inicial de las listas existentes[20] creadas a partir de los trabajos previos del Proyecto, que se referencian como palabras clave a detectar en el texto.

Para avanzar con nuestro objetivo principal de lograr la georreferenciación se dejaron de lado en esta etapa las debilidades de la técnica utilizada. Entre estas debilidades, la dependencia de las diferentes posibilidades de escritura de una palabra, de la ampliación de la lista de palabras clave inicial y de lograr la identificación separada de conocimientos y competencias y su vinculación con los perfiles profesionales. Otra cuestión que podrá ser tenida en cuenta, es la de mejorar el método de clasificación eliminando inicialmente artículos, preposiciones, verbos y otras palabras que no serían útiles en la comparación, o analizando la existencia de corpus específicos asociados a librerías como Natural LanguageToolkit [21] para Python.

3. Desarrollo de Funcionalidades

Todos los programas (scripts) requeridos se construyeron utilizando Python [22], en su versión 2.7.13, como lenguaje de programación. Esta decisión se basó fundamentalmente en el dinamismo, rapidez, facilidad de uso y disponibilidad de librerías de código abierto que Python ofrece para estas tareas, sumado a la experiencia previa que algunos integrantes del equipo contaban con el lenguaje. Todo el código generado se puso bajo control de versiones utilizando Git [23], en un repositorio por el momento privado, aunque no se descarta abrirlo a la comunidad una vez que las herramientas generadas estén lo suficientemente estables y maduras.

3.1. Aplicación de Técnicas de Recolección de Datos

Utilizando las funcionalidades desarrolladas para la recolección de datos, se procedió a la extracción de una muestra cercana a los dos meses de registros. De los datos que inicialmente se obtuvieron, se tomaron como parte de este trabajo:

- **Región**: indica la ciudad y provincia del aviso de oferta laboral.
- **Título**: refiere al texto inicial del aviso que se publica y a partir del cual se accede al cuerpo.
- **Cuerpo**: es el cuerpo completo del aviso.

Se aplicaron luego las técnicas individualizadas para el pre-procesamiento de los datos y extracción de las palabras clave a considerar en el estudio, para posteriormente realizar la localización requerida y la aplicación en un SIG.

3.2. Pre-Procesamiento de Datos

El pre-procesamiento de los datos consistió en la eliminación de avisos duplicados, eliminación o corrección de registros con alguna variable faltante (existieron casos de registros sin región o sin cuerpo de aviso), unificación de regiones, como así también la corrección o eliminación de caracteres especiales o escapes de HTML. Se eliminaron, además, aquellos avisos que no poseían relación alguna con la búsqueda de perfiles orientados a las Tecnologías de la Información.

A partir del cuerpo completo del aviso se obtuvieron las palabras clave (keywords) más relevantes.

Los datos extraídos de cada plataforma se almacenaron en archivos CSV separados por plataforma y por fecha de extracción, para luego ser leídos y unificados por otro script encargado de la integración de los datos y la búsqueda de palabras clave. Para todo ello, fueron de gran utilidad las librerías de Python: **pandas** [24], **json** (para la lectura y escritura de archivos JSON) y **re** (para el manejo de expresiones regulares en la búsqueda de palabras clave).

En la Tabla 1 se visualiza una pequeña muestra de los resultados obtenidos. Se eliminaron algunas columnas de la tabla original por cuestiones de relevancia, claridad y espacio. Por las mismas razones se muestra el título del aviso en lugar del cuerpo completo.

A pesar de que las primeras aproximaciones para la extracción de palabras clave de un aviso de oferta laboral fueron bastante simples y primitivas, sirvieron como base para comenzar a experimentar e ir encontrando fortalezas y debilidades en los métodos utilizados, los cuales fueron evolucionando y refinando continuamente para aumentar su eficacia y eficiencia.

La lista de palabras clave es un archivo JSON2 con objetos y listas específicas de posibilidades. Cuenta actualmente con casi 400 tecnologías a partir de las cuales los hijos en el formato JSON representan las diferentes formas en que se encontraron inicialmente escritas en los avisos tomados como referencia.

Para explicar brevemente, el algoritmo comienza recorriendo la lista de manera secuencial, y en algún punto llega al objeto con clave “ECMAScript” (Ver ejemplo en la Figura 2).

Allí analiza si el listado de hijos contiene elementos. En caso de no contener ningún elemento, se dirige al objeto “default” y busca las palabras clave contenidas en su lista “keywords”. Esta búsqueda se realiza por subcadena en casos de que “strict” sea False, y por expresiones regulares en caso contrario. En el caso analizado, “ECMAScript”

```
"ECMAScript": {
  "hijos": [
    {
      "nombre": "ECMAScript5",
      "keywords": ["ecmascript5",
                   "ecma script 5",
                   "es5"
      ]
    },
    {
      "nombre": "ECMAScript6",
      "keywords": ["ecmascript6",
                   "ecma script 6",
                   "es6"
      ]
    }
  ],
  "default": {
    "nombre": "ECMAScript",
    "keywords": ["ecmascript"],
    "strict": false
  }
},
```

Figura 2. Extracto del archivo de palabras clave en formato JSON.

contiene hijos, por lo que el algoritmo los recorre secuencialmente, buscando como subcadenas las keywords de cada hijo. En caso de encontrar esta keyword en el aviso, no se continúa con el resto (si se encuentra “ecmascript5” no se continúa con “ecma script 5”) y se agrega el valor del campo “nombre” como una keyword asociada al aviso. Sólo cuando se han recorrido todos los hijos sin encontrar coincidencias, se dirige al objeto default para hacer un último intento en encontrar la palabra clave.

Este método de búsqueda, con muchas semejanzas a búsquedas en árboles, permite organizar de manera clara las tecnologías, sus hijos y las palabras claves asociadas a cada uno. Se unifica totalmente el nombre que se coloca en el listado de palabras clave de un aviso; no importa si el aviso dice “ecmascript5”, “ecma script 5”, o “es5”, el listado final contendrá la palabra “ECMAScript5”. Adicionalmente, se puede señalar aquellas palabras clave donde se precisa una búsqueda con expresiones regulares y no como sub-cadena. Un ejemplo de esto es el lenguaje de programación “C”, cuyo objeto “default” tiene como “keywords” a la letra “c” y “strict” está en True, por lo que se inyecta la letra c en una expresión regular que coincide en aquellos casos donde “c” no esté precedida por caracteres alfanuméricos y no tenga delante letras o caracteres especiales (evitando así los casos de C# o C++).

3.3. Georreferenciación de Entidades y Atributos

En función de lo especificado en el objetivo de este trabajo se desarrolla un primer prototipo del SIG utilizando los datos obtenidos como prueba para el mismo.

Cuadro 1. Muestra de los Resultados de la extracción de avisos de ofertas laborales.

Región	Título	Keywords
caba, buenos aires	analista en networking	Web Application Firewall, Networking, Proxys, Seguridad web
caba, buenos aires	analista QA funcional	Testing funcional, Automatización de testing, Analista QA funcional, PHP, Selenium, Casos de prueba
caba, buenos aires	especialista en seguridad informática	OAuth, Seguridad web, Seguridad Informática, Portugués, Web Application Firewall, SAML

En la primera capa se incluyen las ofertas laborales obtenidas a partir de las coordenadas que ubican cada una de los avisos como entidades a las cuales se asocian los perfiles, conocimientos y competencias enunciadas en los mismos como atributos temáticos.

Para lograr la georreferenciación de los avisos y palabras clave, se hace uso de la región en la que se ofrece el puesto laboral. Partiendo de la región se obtienen sus coordenadas geográficas (latitud y longitud) para localizarlas posteriormente en el mapa. Entonces, a los datos que inicialmente se definieron a utilizar, se agregaron:

- **Latitud**: es la latitud de las coordenadas en que se ubica la región.
- **Longitud**: es la longitud de las coordenadas en que se ubica la región.

Habiendo obtenido las palabras clave de cada aviso en particular, se definió inicialmente mostrar en el mapa todas las regiones donde se habían publicado avisos y en las mismas destacar las cinco palabras clave (perfiles, competencias o tecnologías) más mencionadas en ella.

El primer paso consistió en tomar el conjunto de todas las regiones en que se publicó al menos un aviso, e intentar obtener para cada una de ellas su ubicación geográfica aproximada, dada por su latitud y longitud. Para ello se construyó un script, que dada una región (por ejemplo: 'caba, buenos aires' o 'rosario, santa fe') obtiene la latitud y longitud correspondiente consultando a una API externa a través de la librería GeoPy[2]. Para no sobrecargar a la API con una gran cantidad de consultas sucesivas (actualmente se cuenta con más de 7000 registros para georreferenciar), se construyó un archivo JSON (ver Figura 3) que actúa como una cache de regiones, en el cual cada vez que se consulta una región no conocida (que no está en dicho archivo), se la agrega con su denominación y su latitud y longitud asociadas. Además de evitar errores de conexión con las APIs externas por sobrecargas, esta cache de localizaciones agiliza notoriamente el proceso.

[2]Disponible en https://github.com/geopy/geopy


```
{
  "caba, buenos aires": {
    "latitud": -34.5682239,
    "longitud": -58.4479097
  },
  "quilmes, buenos aires": {
    "latitud": -34.7303025,
    "longitud": -58.268868
  },
  "mendoza, mendoza": {
    "latitud": -32.8897294,
    "longitud": -68.8442956
  },
```


Figura 3. Extracto del archivo de regiones, con su latitud y longitud, en formato JSON.

A pesar de la simplicidad que aparenta el proceso una vez resuelto y ejecutado, hay que tener en cuenta que las denominaciones de cada región que se encuentran en los avisos de ofertas laborales son muchas veces ambiguas, contienen errores gramaticales, o no hacen referencia a un lugar específico y conocido por la API consultada. Por ello se debieron tomar las precauciones necesarias para poder identificar regiones por defecto en caso de que una región en particular no fuera encontrada con éxito. En general, se decidió que si una ciudad o distrito no era encontrada pero sí era conocida la provincia a la que pertenecía, entonces se colocaría por defecto dicho aviso en la latitud y longitud de la capital de la provincia. El caso más paradigmático resultó ser la Ciudad Autónoma de Buenos Aires, que se encontró referida de las formas CABA, buenos aires, ciudad de buenos aires o capital federal, lo cual que se unificó en todos los casos a "caba".

El paso siguiente consistió en lograr agrupar, por cada par <latitud, longitud>, el conjunto de palabras clave mencionado y acumular la cantidad de avisos. Fueron de gran ayuda para esta tarea la librería **pandas** junto con las estructuras de diccionarios propias de Python (Dict, OrderedDict, Counter). Una pequeña muestra del resultado de esta etapa se puede apreciar en la Figura 4.

lat	lon	keywords
-34.56	-58.44	Git: 300, SQL Server: 282, Metodologías Ágiles: 187
-37.94	-57.58	SQL: 9, Analista funcional: 8, Desarrollador .NET: 6

Figura 4. Extracto del archivo con frecuencias de palabras clave agrupadas por latitud y longitud.

Esta etapa concluye con la generación de un archivo de texto plano CSV con más de 190 ubicaciones exactas, sus palabras claves asociadas y la frecuencia de aparición de cada una de estas últimas.

Posteriormente, y para verificar la superposición de las capas se ubican, en una segunda capa, las empresas que han participado de la encuesta inicial realizada en la región Rosario[5] como entidades, de forma tal de poder trabajar en un futuro con sectores transversales de aplicación, productos y tecnologías utilizadas. En la tercer y cuarta capa se incluyen las Universidades y Facultades que dictan la carrera de Ingeniería en Sistemas de Información/Informática, inicialmente las incluidas en la Red de Carreras de Ingeniería en Informática / Sistemas de Información del CONFEDI (RIISIC), como entidades, para así trabajar el atributo: oferta de formación.

Del trabajo realizado se ha conseguido generar un diccionario de vértices como avance hacia la definición de la estructura topológica que permitirá el análisis de redes de influencia.

3.4. Aplicación del SIG

El primer paso para la construcción del mapa fue la creación de su capa base, que contiene únicamente el mapa de fondo. Ésta fue creada a través del plugin OpenLayers[3], seleccionando la capa Google Physical. Sobre dicha capa base se agregó luego una nueva capa, que utiliza como fuente de datos el archivo CSV que se mostró en la Figura 4, y permite visualizar todos los puntos geográficos para los cuales se tienen datos registrados.

Una vez importada dicha tabla con las columnas **Latitud, Longitud** y **Keywords**, se procedió a pre-procesar los datos para lograr obtener por cada ubicación, las cinco palabras clave más relevantes.

Listado Código Fuente 1. Función aplicada a cada lista de palabras claves

```
@qgsfunction(args='auto', group='Custom')
def format_text(value, feature, parent):
  value = (value.lstrip('{')
                .rstrip('}'))
  return '<br/>'.join(value.split(',')[:5])
```

[3]Disponible en https://github.com/sourcepole/qgis-openlayers-plugin

En el listado de código fuente 1, se eliminan primero los caracteres { y } del comienzo y final de la lista. Luego, se obtienen las cinco primeras palabras clave (ya vienen ordenados por su frecuencia de aparición, de mayor a menor) y luego se las separa a través del caracter HTML
 por motivos estéticos de visualización en el mapa web, se buscó mostrar la lista en un pequeño pop-up cuando el usuario pasa el cursor por una ubicación marcada (ver Figura 5).

Ya construido el mapa, se añadieron las restantes capas previstas para esta etapa: Empresas y Facultades a las que se agregó como información adicional la dirección de su página web institucional.

El paso final fue exportar el mapa y publicar una primera versión en una web abierta al público. Para ello, fue de gran utilidad el plugin qgis2web[4], que se encarga de convertir, automáticamente y en un único paso, el conjunto de capas y datos del mapa en un conjunto de archivos HTML, Javascript y CSS para su posterior puesta en producción en un servidor web[5].

Las posibilidades que da la herramienta permiten la superposición de todas las capas como se muestra en la Figura 6, o la selección de algunas de las capas y la expansión del mapa como se muestra en la Figura 7).

3.5. Primeros Resultados

Como puede verse en las Figuras 6 y 7, la superposición de las capas inicialmente planteadas 6 permite reconocer áreas de influencia de la trama productiva SSI en nuestro país, en función de las demandas, así como cierta relación con la oferta académica universitaria existente.

A partir de este primer análisis, la inclusión de los restantes atributos a las capas de Empresas y Universidades y Facultades, así como el resto de las capas previstas, como por ejemplo, la capa de profesionales de las diferentes carreras y su ubicación actual, y de la obtención de la topología para el análisis, se podrán identificar, entre otros, las externalidades positivas o negativas que se derivan, o la migración de profesionales a regiones donde la oferta laboral lo justifique.

4. Conclusiones

El trabajo desarrollado ha permitido el descubrimiento de nuevas fortalezas y de debilidades a mitigar para cumplir con los objetivos de IISI.d.r.O. Como se puede apreciar, la mayor dificultad para el desarrollo de este tipo de sistemas radica en la definición del enfoque y objetivos del análisis que se desea realizar, junto con obtención de los datos a trabajar, más que en la aplicación de las técnicas de georreferenciación.

[4]Disponible en https://github.com/tomchadwin/qgis2web

[5]Versión inicial del mapa en http://mapa-iisidro.000webhostapp.com

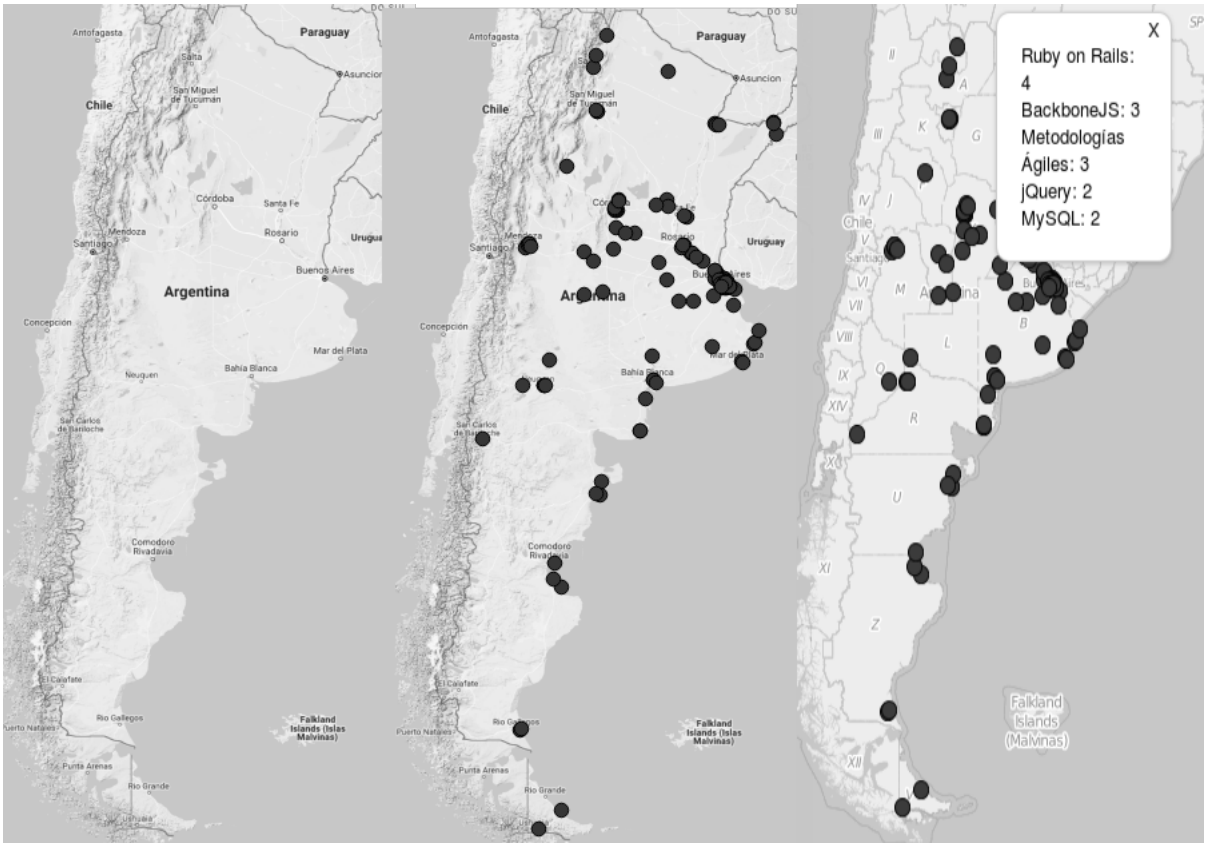

Figura 5. Versión inicial del mapa: capa base, ubicaciones y mapa exportado a web

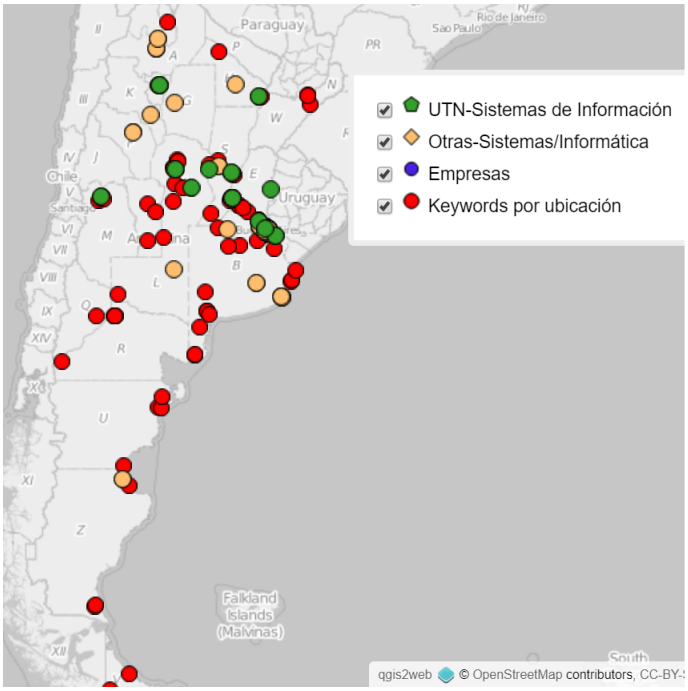

Figura 6. Mapa final obtenido producto del trabajo

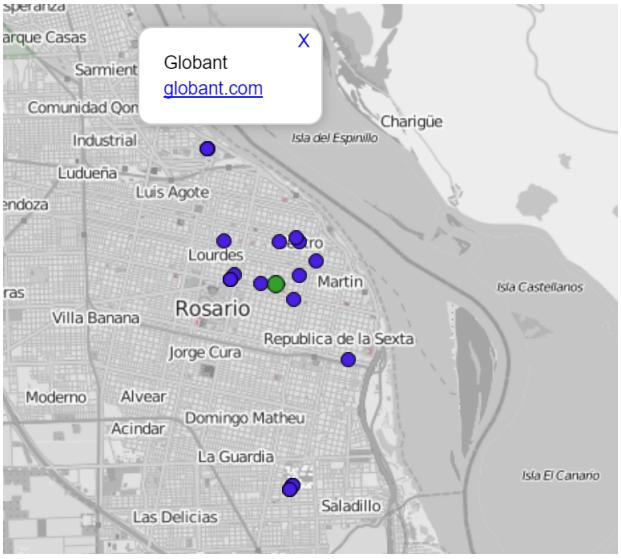

Figura 7. Mapa Región Rosario Capas: Empresas-UTN

En cuanto a la obtención de datos, si bien la solución implementada otorga excelentes resultados y en tiempos de ejecución casi instantáneos para un conjunto cercano a los siete mil avisos, se han identificado, al mencionar las técnicas utilizadas, cuestiones en las que avanzar para lograr una mayor efectividad. Sin embargo, del trabajo realizado ya se puede pensar en la aplicación de técnicas más avanzadas de extracción de conocimiento y aprendizaje automáticos.

En cuanto a la aplicación en el SIG, el hecho de construir el mapa a través de herramientas como QGIS brinda una gran flexibilidad a la hora de la presentación de los datos para el usuario, dando a su vez importantes posibilidades para, en un futuro, agregar sucesivas capas con datos de otras fuentes y así escalar la solución sin mayores inconvenientes. Resta avanzar sobre la posibilidad de realizar un estudio topológico en función de la información obtenida, buscando ofrecer una herramienta útil para la Gestión del Conocimiento en el ámbito de IISI.d.r.O. como aporte a los sectores que conforman el triángulo de Sábato: Universidad-Estado-Industria.

Referencias

[1] F. M. Riva, E. Amar, V. Martín, M. A. Gatto, y N. Pereira, "Observatorio de desarrollo regional de la ingeniería en sistemas de información e informática (IISI.d.r.O.). Origen, Evolución y Perspectivas.," *En Memorias: CONAIISI 2016. IV Congreso Nacional de Informática e Ingeniería en Sistemas de Información. UCASAL. Publicación on line - ISSN 2347-0372*, 2016.

[2] R. G. Malano, V. A. Martin, y F. M. Riva, "Favoreciendo el desarrollo de conocimientos y competencias en el contexto de un proyecto de investigación," *Iberoamerican Journal of Project Management*, vol. 8, no. 1, pp. 01–13, 2017.

[3] F. M. Riva, E. Amar, V. Martín, y N. Pereira, "Una red para el análisis comparado de competencias en la trama productiva de la industria del software y servicios informáticos," *En revista: Rumbos Tecnológicos de la Secretaría de Ciencia, Tecnología y Posgrado de la UTN-FRA.*, vol. 8, pp. 135–143, 2016.

[4] F. M. Riva y M. Kain, "Informe técnico 2: Requerimientos de RRHH de empresas cadena productiva ssi," *PID UTN-1923: Modelización de un Observatorio de Desarrollo Productivo. Industria del Software y Servicios Informáticos en el Área Rosario.*, 2014. Disponible en: `http://isi-investiga.frro.utn.edu.ar/observatorio`. [Última fecha de acceso: 01/08/2017].

[5] F. M. Riva, E. Amar, V. Martín, E. Porta, C. Galmarini, y M. Puyo, "Informe técnico 1: Relevamiento a empresas del sector ssi," *PID UTN-1923: Modelización de un Observatorio de Desarrollo Productivo. Industria del Software y Servicios Informáticos en el Área Rosario.*, 2014. Disponible en: Disponible en: `http://isi-investiga.frro.utn.edu.ar/observatorio`. [Última fecha de acceso: 01/08/2017].

[6] F. M. Riva, E. Amar, V. Martín, y N. Pereira, "Informe técnico 2: : Encuestas a egresados y estudiantes que desarrollan su actividad en la cadena productiva ssi.," *PID UTN-1923: Modelización de un Observatorio de*

Desarrollo Productivo. Industria del Software y Servicios Informáticos en el Área Rosario., 2016. Disponible en: http://isi-investiga.frro.utn.edu.ar/observatorio. [Última fecha de acceso: 01/08/2017].

[7] J. M. Rodriguez Guerrero, M. Abbatemarco, y J. García, "Herramientas para la implementación de una red orientada al análisis de competencias de egreso de la carrera de ingeniería en sistemas de información.," *JIT 2017. Jornadas Investigadores Tecnológicos. Reconquista, Santa Fe, Argentina. Seleccionado para su publicación en la Revista Tecnología y Ciencia de la Secretaría de Ciencia, Tecnología y Posgrado de UTN.*, 2017.

[8] M. Abbatemarco, L. Brizuela, A. Cervino, y F. M. Riva, "Las redes sociales como fuente de datos para un observatorio regional de ingeniería en sistemas de información e informática. oportunidades y limitaciones técnicas, éticas y legales.," *Memorias: CONAIISI 2016. IV Congreso Nacional de Informática e Ingeniería en Sistemas de Información. UCASAL. Publicación on line - ISSN 2347-0372*, 2016.

[9] F. M. Riva, M. Abbatemarco, y A. Cervino, "El tratamiento masivo de datos en redes sociales virtuales. retos legales, éticos y de responsabilidad social," *En Memorias de XLIII CLEI - 46 JAIIO. Jornadas Argentinas de Informática. Córdoba. Argentina. Publicación on line: ISSN 1850-2776*, 2017.

[10] M. Fuenzalida, G. D. Buzai, A. Moreno Jiménez, y A. García de León, *Geografía, Geotecnología y Análisis Espacial: Tendencias, Métodos y Aplicaciones.* Editorial Triángulo. Santiago de Chile., 2013.

[11] L. M. Humacata, "Aportes metodológicos del análisis espacial con sistemas de información geográfica a la clasificación espacial en geografía," *En Revista: Redes Sociales*, vol. 3, pp. 118–147, 2014.

[12] G. D. Buzai y C. A. Baxendale, "Análisis socioespacial con sistemas de información geográfica marco conceptual basado en la teoría de la geografía," *Ciencias Espaciales*, vol. 8, no. 2, pp. 391–408, 2015.

[13] G. D. Buzai, "Geografía aplicada a la solución de problemáticas sociales," *SOLUCIONES ESPACIALES A PROBLEMAS SOCIALES URBANOS*, p. 17, 2016.

[14] I. del Bosque González, C. F. Freire, L. M.-F. Morente, y E. P. Asensio, "Los sistemas de información geográfica y la investigación en ciencias humanas y sociales," 2012.

[15] Quantum GIS Development Team, "Quantum GIS geographic information system. Open Source Geospatial Foundation project." http://qgis.osgeo.org, 2017.

[16] E. Ferrara, P. De Meo, G. Fiumara, y R. Baumgartner, "Web data extraction, applications and techniques: A survey," *Knowledge-based systems*, vol. 70, pp. 301–323, 2014.

[17] A. H. Laender, B. A. Ribeiro-Neto, A. S. Da Silva, y J. S. Teixeira, "A brief survey of web data extraction tools," *ACM Sigmod Record*, vol. 31, no. 2, pp. 84–93, 2002.

[18] K. Devika y S. Surendran, "An overview of web data extraction techniques," *International journal of scientific engineering and technology*, vol. 2, no. 4, 2013.

[19] I. H. Witten, E. Frank, M. A. Hall, y C. J. Pal, *Data Mining: Practical machine learning tools and techniques*. Elsevier. Morgan Kaufmann Publishers, 2016.

[20] G. de la Calle Velasco, *Modelo basado en técnicas de procesamiento de lenguaje natural para extraer y anotar información de publicaciones científicas*. PhD thesis, Tesis Doctoral. Escuela Técnica Superior de Ingenieros Informáticos. Universidad Politécnica de Madrid., 2014.

[21] S. Bird, "Nltk: the natural language toolkit," in *Proceedings of the COLING/ACL on Interactive presentation sessions*, pp. 69–72, Association for Computational Linguistics, 2006.

[22] Python Software Foundation, "Python language reference, versión 2.7.13." https://www.python.org, 2017.

[23] Git, "Sistema de control de versiones distribuido de código abierto." https://git-scm.com, 2017.

[24] W. McKinney, "Data structures for statistical computing in Python," in *Proceedings of the 9th Python in Science Conference*, pp. 51–56, 2010.