Extracción y Normalización Automática de Información en la Búsqueda de Componentes SIG

Gabriela Gaetán[1], Agustina Buccella[2], Alejandra Cechich[2]

[1]Proyecto de Investigación Área Ingeniería de Software
Unidad Académica Caleta Olivia – Universidad Nacional de la Patagonia Austral
ggaetan@uaco.unpa.edu.ar
[2]Grupo de Investigación en Ingeniería de Software del Comahue (GIISCo)
Departamento de Ciencias de la Computación
Universidad Nacional del Comahue
{abuccel, acechich}@uncoma.edu.ar

Resumen. Uno de los problemas en el crecimiento del desarrollo de software basado en componentes es la dificultad para localizar y recuperar componentes de software existentes. Las características particulares de los Sistemas de Información Geográficos (SIG) provocan que las composiciones de componentes no puedan construirse sólo como simples piezas de un producto, sino como un conjunto de elementos pre-ensamblados lo que hace la identificación de candidatos particularmente compleja. En ese contexto, este artículo presenta un proceso para publicar información sobre componentes SIG almacenada en repositorios Web, estructurándola en base a un esquema de información normalizado y enriquecido por medio de técnicas de Procesamiento del Lenguaje Natural. Se describen los principales elementos de la herramienta que automatiza este proceso y se evalúan los resultados experimentales de un caso de estudio.

Palabras claves: DSBC, OTS, servicios SIG, ontologías, lenguaje natural.

1 Introducción

Con el surgimiento del desarrollo de software basado en componentes, numerosas empresas fabricantes de SIG han comenzado a comercializar distintos tipos de componentes software orientados a las necesidades de los desarrolladores SIG. Para lograr un desarrollo más eficiente, los analistas se concentran en los atributos de reusabilidad e interoperabilidad. Sin embargo, se pierde mucho tiempo y esfuerzo en encontrar aquellos componentes que satisfagan la funcionalidad que se pretende implementar. Una de las necesidades clave para facilitar esta tarea, consiste en contar con información estándar de los componentes que permita agilizar la búsqueda de composiciones de software.

Este trabajo se presenta como una extensión a los trabajos presentados en [5] y [6] en los cuales hemos propuesto un Proceso de Publicación de componentes SIG basado en un Esquema de Clasificación normalizado aplicando técnicas de procesamiento de lenguaje natural.

En la literatura se encuentran distintos aportes relacionados con la publicación de componentes [2]. Algunos trabajos, [1], [4] y [8], proponen resolver los problemas relacionados con la categorización (o indexación), concentrando sus propuestas principalmente en las estructuras de clasificación de la información. Otros, como Componex[1] y ComponentXchange [9], además avanzan en la solución de los problemas de búsqueda y almacenamiento de información. Mientras que otras propuestas, como en [7], introducen técnicas de procesamiento de lenguaje natural, ontologías y modelos de dominio, para incrementar la efectividad de la búsqueda.

A diferencia de estos trabajos, nuestra propuesta se concentra en las ventajas que el conocimiento de un dominio específico aporta a la selección de componentes. De esa manera, estándares del dominio geográfico se usan para normalizar la información y permitir el modelado de una base conceptual que derive en una búsqueda automática. La contribución de este artículo se centra precisamente en la presentación de un prototipo que recibiendo información de catálogos de componentes SIG publicados en la Web, normaliza y clasifica la información de manera automática.

Este artículo está organizado de la siguiente manera. A continuación en la Sección 2, se describe brevemente el proceso así como la herramienta de soporte para extracción y normalización automática. En la Sección 3 se muestra una primera evaluación de la herramienta. Finalmente en la última sección se presentan las conclusiones y trabajo futuro.

2 Proceso de Publicación de Componentes SIG

En trabajos previos [5] y [6] hemos propuesto un Proceso de Publicación de componentes SIG, basado en un Esquema de Clasificación normalizado que considera información funcional y no técnica de componentes SIG. La Figura 1 representa esquemáticamente este proceso. La Ontología que usa el Proceso de Publicación es generada por el módulo Creación de Ontología. Este proceso es asistido por un experto humano y se implementa por medio de un editor de ontologías como Protégé[2]. El Usuario "Publicador" procesa una Descripción web (que representa la información sobre los componentes encontrada en los catálogos) por medio del módulo Extracción de Información. Este módulo aplica técnicas de anotación semántica para lo cual se utiliza la Ontología generada. El módulo Normalización de información estructura el documento que contiene la descripción web en forma de una Descripción normalizada. Antes de almacenar definitivamente la información normalizada en el Repositorio Estandarizado, también se evalúa la validez de las anotaciones obtenidas automáticamente.

[1] http://www.componex.biz
[2] http://protege.stanford.edu/

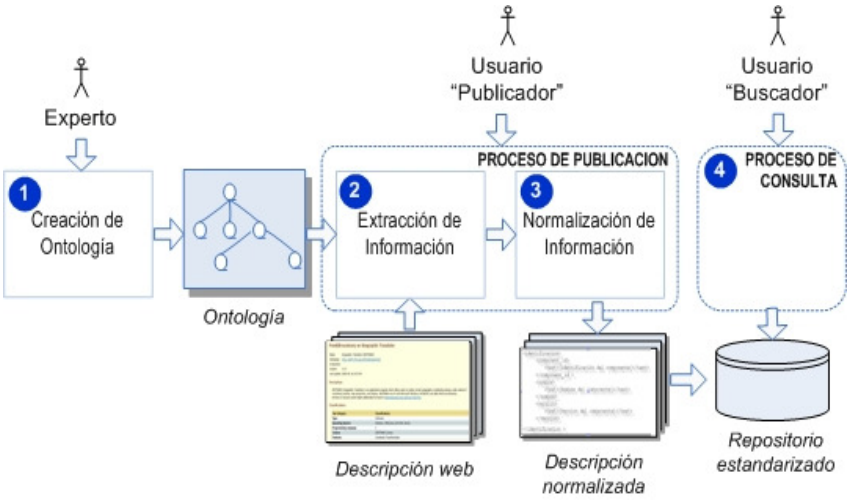

Figura 1. Partes principales del proceso de Publicación de Componentes

El proceso de Publicación se puede relacionar con otro proceso externo: Proceso de Consulta, mediante el cual, un Usuario ("Buscador") que busca información sobre componentes podría hacer uso de la información extraída y almacenada en el Repositorio estandarizado.

Creación de la Ontología: Este módulo se encarga de la implementación de la ontología ontoCompoSIG [5], para ello se utiliza el editor de ontologías Protégé v4.1.0. y se genera una ontología en formato OWL-DL. Como punto de partida usamos una ontología disponible (ISO 19119 Service Type[3] de GEOBRAIN) que representa la taxonomía de servicios geográficos del estándar ISO/IEC19119. El modelo conceptual de la ontología se basa principalmente en el Esquema de clasificación normalizado [6], que pretende organizar en forma estructurada la información que describe a los componentes SIG. Estos elementos conceptuales se implementan en tres clases principales:

- Funcionalidad: esta clase representa los aspectos funcionales de los componentes SIG, y está formada por dos sub-clases: (1) Tarea – la tarea que realiza el componente; y (2) Datos – representa todos los objetos que pueden ser considerados como entradas o salidas de las tareas que realiza el componente.
- Descripción: esta clase representa los datos particulares que describen a los componentes SIG. Se organiza en las siguientes subclases: ComponenteSIG, Contacto, RequerimientoSoftware, RequerimientoHardware, Artefacto, SitioWeb.
- Clasificación: esta clase representa facetas de clasificación de los componentes SIG. Se organiza en las siguientes subclases: Estandar, SistemaOperativo, Aplicacion, EstadoDesarrollo, Idioma, Licencia, LenguajeProgramacion, ServicioGeografico (las diferentes sub-clases se organizan usando la ontología "ISO 19119 Service Type" de GEOBRAIN).

[3] http://geobrain.laits.gmu.edu/ontology/2004/12/ISO_geographic_service.owl

Extracción de la Información: El primer paso del Proceso de Publicación consiste en extraer toda la información relevante relacionada con el dominio de los componentes SIG desde documentos disponibles en la web. Como se muestra en la Figura 2, el primer paso de Inicialización del texto realiza la preparación del documento a ser procesado liberándolo de cualquier etiqueta que no sea necesaria para el resto del proceso. En el paso de Delimitación del texto se divide al texto en bloques de construcción básica (palabras, sentencias o párrafos) y se generan etiquetas que identifican estos bloques básicos para usos futuros. Para la Identificación de las entidades del Esquema de clasificación normalizado dentro del texto disponible se plantean dos estrategias:

1. Identificación basada en listas enumeradas. Esta identificación se basa en una lista que permite enumerar los valores posibles de un concepto y asociarlos a una ontología. Por ejemplo, la categoría 'Estado de desarrollo' de un componente tiene como valores posibles: 'stable, alfa, beta, archived', y se asocia al concepto 'EstadoDesarrollo' de la ontología OntoCompoSIG.
2. Identificación basada en patrones del contexto. Esta identificación se basa en la aplicación de reglas gramaticales solo si ocurre una cierta situación en el texto. Por ejemplo, se define una regla para la categoría 'Nombre de Componente', que indica que el nombre de un componente solo puede ser reconocido si, en el texto, éste ocurre precedido por la palabra 'Name'. En este paso también sería posible aplicar reglas que permitan inferir conocimiento no declarado explícitamente en el documento original.

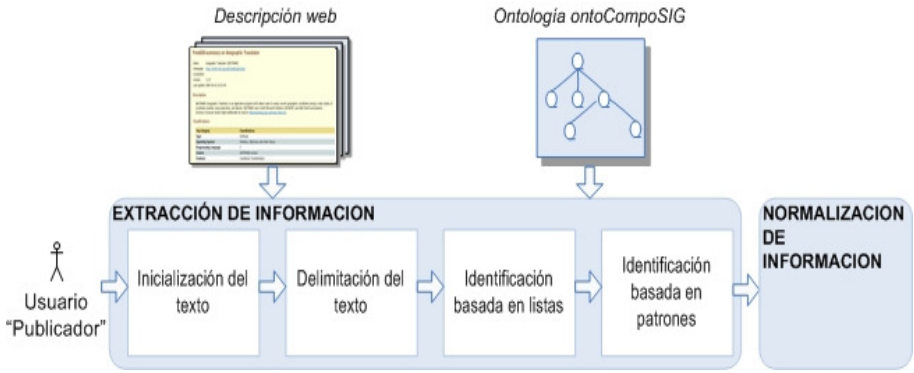

Figura 2. Detalle del Módulo 2: Extracción

El resultado final del proceso de Extracción de información está formado por el texto (encontrado en el documento original) más un conjunto de anotaciones conceptuales que enlazan partes de ese texto con conceptos de la ontología específica del dominio SIG (ontoCompoSIG).

Normalización de la Información: Como se muestra en la Figura 3, este módulo está formado por dos pasos secuenciales: Evaluación de anotaciones y Estructuración de información.

Figura 3. Detalle del Módulo 3: Normalización

En el paso de Evaluación de anotaciones se pretende garantizar la completitud y correctitud de la información que se almacenará en el Repositorio estándar. El usuario interviene para analizar las anotaciones automáticas que realizó el módulo Extracción de Información, de modo que sea posible completar correctamente la máxima cantidad de elementos del Esquema de clasificación normalizado.

Durante la Estructuración de información se organiza la información disponible (el contenido original del documento + las anotaciones realizadas) para dar forma a un documento que contenga la Descripción normalizada. Esta información, que se almacenará en el Repositorio estandarizado, consiste en un conjunto de meta-datos que permiten completar el Esquema de clasificación normalizado. Luego, otros usuarios pueden acceder a este repositorio para encontrar y recuperar aquellos componentes que más se adecuen a sus necesidades (Módulo 4 – Figura 1).

2.1 Herramienta de Soporte para la Extracción de Información

Existen numerosas herramientas que brindan soporte para desarrollar aplicaciones en el dominio del Procesamiento de Lenguaje Natural. Para este trabajo, se eligió GATE (General Architecture for Text Engineering) porque, según se sostiene en [3], ofrece soporte para aplicar técnicas de procesamiento de lenguaje basadas en ontologías. GATE está basada en Java y está disponible bajo licencia LGPL (Lesser GNU Public Licence). JAPE (Java Annotation Patterns Engine) es un módulo de GATE que permite realizar transformaciones en las anotaciones de un texto previamente anotado. ANNIE (a Nearly-New Information Extraction System) es un sistema de extracción de información que se distribuye con GATE y usa técnicas de máquinas de estado finito para implementar tareas como tokenización, etiquetado semántico y particionamiento de frases. Los pasos del módulo Extracción de Información, que se muestran en la Figura 4, se implementaron completamente desarrollando una aplicación conformada por una secuencia de componentes disponibles en GATE.

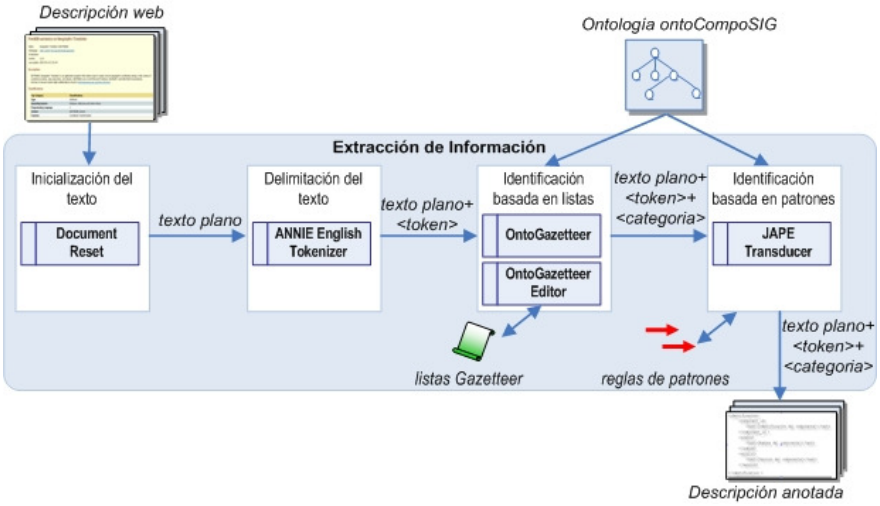

Figura 4. Implementación del módulo Extracción de Información.

3 Evaluación Experimental

En nuestro escenario, el usuario selecciona un documento desde un repositorio local, que contiene un conjunto de documentos con información de los componentes SIG ofrecidos en el catálogo FreeGIS[4]. Para este ejemplo seleccionaremos la descripción del componente "Geographic Translator" que se muestra en la Figura 5.

FreeGIS-summary on Geographic Translator

Name: Geographic Translator (GEOTRANS)
Homepage: http://earth-info.nga.mil/GandG/geotrans/
Screenshot:
Version: 2.2.5
Last update: 2005-06-16 23:52:40

Description:

GEOTRANS (Geographic Translator) is an application program which allows users to easily convert geographic coordinates among a wide variety of coordinate systems, map projections, and datums. GEOTRANS runs in both Microsoft Windows (95/98/NT) and UNIX Motif environments. Archives of several version might additionally be found in Remotesensing.org's geotrans directory.

Classifications:

Top Category	Classifications
Type	Software
Operating System	Windows, GNU/Linux and other Unices
Programming Language	C
License	GEOTRANS License
Features	Coordinate Transformation

Links:

No links have been defined.

Changes:

Figura 5. Descripción de un componente en lenguaje natural presentada en un catálogo web.

[4] http://freegis.org/

Escenario de uso:

1. El usuario selecciona la ontología ontoCompoSIG e inicia la aplicación 'Extracción de información' para el documento seleccionado.
2. El prototipo remueve las anotaciones previas y cualquier marca (markup) no deseada del documento seleccionado. En este caso, se eliminan las etiquetas de marcado HTML que contiene el documento original.
3. El prototipo divide el texto en tokens y crea anotaciones Token en el documento. Luego identifica los términos relacionados con las clases de la ontología: 'Sistema Operativo', 'Licencia', 'Lenguaje de Programación', 'Tarea' y 'Datos'. Crea anotaciones Lookup con el atributo 'class' relacionado con la clase correspondiente de la ontología. Finalmente, identifica otros detalles como 'Nombre', 'Version' y 'Sitio web' en base a una serie de reglas gramaticales basadas en el contexto. Crea anotaciones Lookup con el atributo 'class' relacionado con la clase correspondiente de la ontología.
4. El prototipo permite que el usuario (1) evalúe el resultado del proceso de extracción automático y (2) pueda hacer los cambios que considere necesarios por medio de anotación manual.
5. Finalmente, el prototipo recolecta solamente las anotaciones importantes realizadas sobre el documento aplicando un conjunto de reglas gramaticales.
6. El usuario almacena el documento resultante como un documento XML, en el Repositorio estandarizado.
7. Como resultado del procesamiento antes detallado, en la Tabla 1 se muestran los términos identificados para la descripción web de un componente.

Tabla 1. Información de Geographic Translator relacionada con la ontología.

Concepto	***Descripción del componente SIG***
Nombre	GeographicTranslator
Version	2.2.5
sitioWeb	http://earth-info.nga.mil/GandG/geotrans
SistemaOperativo	Windows, GNU/Linux, other Unices
lenguajeProgramacion	C
Licencia	GEOTRANSLicense
Datos	Geographic Coordinates, Map projections
Tarea	Convert

3.1 Evaluación de la Herramienta

Los documentos sobre los que se realizó esta evaluación experimental se obtuvieron a partir del catálogo web disponible en FreeGIS. Cada documento del catálogo contiene información semi-estructurada sobre un componente SIG, y está escrito en lenguaje natural, en idioma inglés.

Para el proceso de evaluación se seleccionaron aleatoriamente 50 documentos de aproximadamente 50 kb. de tamaño cada uno, que fueron accedidos manualmente por medio de un navegador web y almacenados con formato HTML en un repositorio local. Este conjunto de documentos conforman lo que se conoce como Corpus; a su

vez, este conjunto total se dividió en dos corpus distintos utilizados en las distintas iteraciones del proceso:

- Corpus-1: formado por un conjunto de 10 documentos, utilizados en la Iteración-1 e Iteración-2.
- Corpus-2: formado por los 40 documentos restantes, utilizados en la Iteración-3.

Para poder comparar la calidad de las anotaciones realizadas por nuestro prototipo hemos utilizado las métricas de *Precisión*, *Cobertura* (Recall) y *Medida-F* (F-measure). La Precisión mide el número de entidades identificadas correctamente como un porcentaje de todas las entidades identificadas. Se define como:

$$\text{Precisión} = (\text{Correctas} + 1/2\text{Parciales}) /(\text{Correctas} + \text{Falsas} + \text{Parciales})$$

La Cobertura mide el número de entidades identificadas como un porcentaje de las entidades correctas. Cuanto más alto es el índice de cobertura, mas asegura el sistema que no pierde entidades correctas. Se define como:

$$\text{Cobertura} = (\text{Correctas} + 1/2\text{Parciales}) /(\text{Correctas} + \text{Perdidas} + \text{Parciales})$$

La Medida-F combina Cobertura y Precisión en una única medida. Se define como:

$$\text{F-measure} = ((\beta^2 + 1)\ \text{Precisión}*\text{Cobertura}) / ((\beta^2\ \text{Precisión}) + \text{Cobertura})$$

donde β es un factor que indica la importancia relativa de Cobertura y Precisión.

En las fórmulas, *Falsas* representa a aquellas entidades que son anotadas como correctas, pero no lo son; y *Perdidas* a aquellas entidades que no son encontradas por medio del procesamiento de anotación automático.

Para evaluar la performance del prototipo es necesario tener un texto de referencia, generalmente conocido como "gold standard", que es una anotación manualmente realizada, y se usa para hacer comparaciones con los resultados derivados del proceso de anotación automático. Entonces, el diseño de este experimento, incluye un proceso de evaluación formado por tres pasos:

1. Anotación manual: El anotador humano realiza las anotaciones manuales sobre cada documento que integra el Corpus-n para crear el corpus gold standard. Se obtiene el Corpus-n-manual.
2. Anotación automática: Sobre los mismos documentos que integran el Corpus-n (original, sin anotaciones) se ejecuta la aplicación 'Extracción de Información' para anotarlo automáticamente. Se obtiene el Corpus-n-automático.
3. Comparación: Cada documento del Corpus-n-automático es analizado con respecto a su similar del Corpus-n-manual para obtener las medidas Precisión y Cobertura.

La herramienta Annotation Diff que ofrece GATE permite comparar un texto anotado automáticamente con un texto de referencia (anotado manualmente) y calcula las métricas Precisión, Cobertura y F-Measure. En nuestro caso, la herramienta Annotation Diff identificó 27 anotaciones Correctas, 1 Parcial, 3 Perdidas y 49 Falsas. Al aplicar las fórmulas para las métricas, obtenemos los siguientes valores:

$$\text{Precisión} = (27 + ½*3) /(27 + 49+ 1)= 0,36$$

$$\text{Cobertura} = (27 + ½*3) /(27 + 3 + 1)= 0,89$$

$$\text{F-measure} = 2\ (0,36*0,89) / (0,36+ 0,89)= 0,47$$

La primera parte de nuestro experimento incluye dos iteraciones en las que se ejecuta la aplicación 'Extracción de Información' sobre el mismo conjunto de pruebas (Corpus-1). Al finalizar la Iteración-1, las anotaciones automáticas se comparan con las anotaciones manuales y los resultados son expresados usando las métricas Precisión y Cobertura. La Tabla 2 muestra estos resultados, tanto para la Iteración-1 como para la Iteración-2, a nivel corpus.

Si observamos los resultados de la Iteración-1, cuyos valores promedio para Cobertura y Precisión son 0,70 y 0,69 respectivamente, vemos que solo 6 documentos tienen ambos valores mayores que 0,5. Al examinar manualmente los documentos procesados detectamos dos fuentes principales de errores:

- Errores de reglas gramaticales definidas en la Identificación basada en patrones: por ejemplo, en el documento 'FreeGIS_org2.htm' la anotación automática no reconoce al texto 'http://www.remotesensing.org/proj' como una entidad 'sitioWeb'. Esto se debe que la regla original sólo reconoce texto que se ajuste a un patrón tipo: {"http://"}{palabra}{punto}{palabra}{punto}{palabra}{simbolo}, adecuado para un texto como: 'http://www.misitio.org/'.
- Vocabulario del dominio utilizado en la Identificación basada en listas incompleto: por ejemplo, el texto 'other Unices' fue identificado en la anotación manual como una entidad 'sistemaOperativo', pero en la definición original del vocabulario que define la entidad no había sido considerado.

Tabla 2. Mediciones de Precisión y Cobertura para la evaluación del prototipo.

	Iteración 1		**Iteración 2**		**Diferencia**	
	Cobertura	Precisión	Cobertura	Precisión	Cobertura	Precisión
FreeGIS_org1.htm	0,89	0,36	1	1	0,11	0,64
FreeGIS_org2.htm	0,42	0,64	0,9	1	0,48	0,36
FreeGIS_org3.htm	0,75	0,64	0,96	0,66	0,21	0,02
FreeGIS_org4.htm	0,26	0,76	1	0,92	0,74	0,16
FreeGIS_org5.htm	0,68	0,55	0,92	0,83	0,24	0,28
FreeGIS_org6.htm	1	0,67	1	0,88	0	0,21
FreeGIS_org7.htm	0,9	0,9	0,9	1	0	0,1
FreeGIS_org8.htm	0,81	0,74	0,92	0,8	0,11	0,06
FreeGIS_org9.htm	0,8	0,8	0,96	0,82	0,16	0,02
FreeGIS_org10.htm	0,49	0,81	0,97	0,81	0,48	0
Promedio	**0,70**	**0,69**	**0,95**	**0,87**	**0,25**	**0,19**

En la segunda iteración, luego de ajustar los errores identificados en la Iteración-1, comparamos los documentos del Corpus-1 anotados manualmente con los documentos anotados con el prototipo mejorado. En la Tabla 2 se observa que los valores de Cobertura y Precisión de la Iteración-2 son mayores o iguales a los de la Iteración-1, para todos los documentos.

Sin embargo, las mejoras se reflejan principalmente en la cobertura, y no tanto en la precisión. En promedio, la Cobertura de la Iteración-2 obtiene un valor de 0,95 y la Precisión de 0,87, mostrando una mejora promedio de 0,25 y 0,19 respectivamente para cada una de las métricas. Esto fue producto de que los ajustes para mejorar la Cobertura fueron pequeños cambios en las reglas definidas y en el vocabulario; pero,

para mejorar la Precisión la tarea no es tan directa, requiriendo una modificación de las reglas y/o del vocabulario personalizado para que el prototipo no realice anotaciones donde no corresponde (generación de anotaciones falsas).

4 Conclusión y Trabajos Futuros

En este trabajo hemos detallado los principales elementos del proceso centrado en la normalización de la información (esquema enriquecido semánticamente) y en la recuperación de información desde catálogos existentes usando técnicas del lenguaje natural. El objetivo es identificar de forma automática, a partir de un documento escrito en lenguaje natural, la mayor cantidad posible de las categorías definidas en el esquema normalizado. La herramienta desarrollada permite automatizar este proceso, aplicado a un conjunto de descripciones de componentes encontradas en portales especializados. La primer evaluación ha mostrado resultados prometedores después del refinamiento realizado. Sin embargo, en esta primera fase, tanto el desarrollo de herramientas, como la experimentación y el análisis de los resultados se relacionaron con el proceso de Publicación de descripciones de componentes, mientras que en una segunda fase también investigaremos los resultados del proceso de Recuperación.

Referencias

1. Ackermann J., Brinkop F., Conrad S., Fettke P., Frick A., Glistau E., Jaekel H., Kotlar O., Loos P., Mrech H., Ortner E., Overhage S., Raape U., Sahm S., Schmietendorf A., Teschke T., Turowski K. Standardized Specification of Business Components. German Society of Informatics, 2002.
2. Cechich A., Réquilé A., Aguirre J., Luzuriaga J. Trends on COTS Component Identification. 5th International Conference on COTS-Based Software Systems. IEEE Computer Science Press, 2006.
3. Cunningham H., Maynard D., Bontcheva K. y Tablan V. GATE: A Framework and Graphical Development Environment for Robust NLP Tools and Application. 40th Anniversary Meeting of the Association for Computational Linguistics, 2002.
4. Dong J., Alencar P. S. C., Cowan D. D. A Component Specification Template for COTS-based Software Development. First Workshop on Ensuring Successful COTS Development, 1999.
5. Gaetán G., Cechich A., Buccella A. Aplicación de técnicas de procesamiento de lenguaje natural y web semántica en la publicación de componentes para SIG, Simposio Argentino de Ingeniería de Software, JAIIO, 2009.
6. Gaetán G., Cechich A., Buccella A., Extracción de información a partir de catálogos web de componentes para SIG, XV Congreso Argentino en Ciencias de la Computación, 2009.
7. Girardi M. R., Ibrahim, B. A software reuse system based on natural language specifications. 5th International Conference on Computing and Information, 1993.
8. Torchiano M., Jaccheri L., Sørensen C., Wang I.: COTS Products Characterization. 14th international conference on Software engineering and knowledge engineering, 2002.
9. Varadarajan, S.; Kumar, A.; Gupta, D.; Jalote, P.: ComponentXchange: An E-Exchange For Software Components. IADIS Conf. WWW/Internet, 2001.