TENDENCIAS EN LA INGENIERÍA DEL LENGUAJE Y DEL CONOCIMIENTO

EDITORES:
MIREYA TOVAR VIDAL
DARNES VILARIÑO AYALA
BEATRIZ BELTRÁN MARTÍNEZ

2016

Tendencias en la Ingeniería del Lenguaje y del Conocimiento

Tendencias en la Ingeniería del Lenguaje y del Conocimiento

Realizado en
Puebla, Pue., México.
Otoño 2016.

Tendencias en la Ingeniería del Lenguaje y del Conocimiento

2015-2016

Editores
Mireya Tovar Vidal
Darnes Vilariño Ayala
Beatriz Beltrán Martinéz

BENEMÉRITA UNIVERSIDAD AUTÓNOMA DE PUEBLA

José Alfonso Esparza Ortíz
Rector

René Valdiviezo Sandoval
Secretario General

María del Carmen Martínez Reyes
Vicerrectora de Docencia

Flavio Marcelino Guzmán Sánchez
Encargado de despacho
Vicerrectoría de Extensión y Difusión de la Cultura

Ygnacio Martínez Laguna
Vicerrector de Investigación y Estudios de Posgrado

Ana María Dolores Huerta Jaramillo
Directora de Fomento Editorial

Marcos González Flores
Director de la Facultad de Ciencias de la Computación

Primera Edición **Otoño 2016**
ISBN: 978-607-525-149-3

Dirección de Fomento Editorial
4 Sur 104, CP. 72000
Puebla, Pue., México
Teléfono y Fax: 01 222 246 85 59

Hecho en México
Made in México

Prólogo

Este libro tiene como intención la divulgación de trabajos iniciales en el área de tecnología del lenguaje y del conocimiento. El desarrollo de espacios de publicación o divulgación son mecanismos importantes que deben ser creados en las unidades acádemicas para difundir los resultados alcanzados tanto por el docente como por el alumno en las diferentes líneas de investigación que se cultivan.

En la primera parte de esta obra se presentan algunos estudios que inician directamente vinculados con el área de Tecnología del Lenguaje. En la segunda parte, se discuten resultados alcanzados en temas de investigación que muestran un nivel de madurez adecuado para este tipo de publicación. Se ha creado un comite evaluador por pares a ciegas de cada uno de los trabajos, que logra que la calidad de la publicación sea adecuada.

Esta publicación es la primera versión de extensiones de trabajos enviados al congreso LKE'2015, que necesitaban nuevos experimentos principalmente para la segunda parte de este libro.

Los editores,
Mireya Tovar Vidal
Darnes Vilariño Ayala
Beatriz Beltrán Martínez

Índice general

Parte I

Investigación en la Ingeniería del Lenguaje y del Conocimiento

Capítulo 1

Descubrimiento y representación de conocimiento sobre datos científicos y académicos

José A. Reyes-Ortiz

Departamento de Sistemas,
Universidad Autónoma Metropolitana, Azcapotzalco
Av. San Pablo 180, Azcapotzalco, 02200
Ciudad de México, México.
jaro@correo.azc.uam.mx

Resumen. La búsqueda de información académica y publicaciones científicas es de suma importancia en los últimos años debido al crecimiento acelerado de los datos disponibles en la Web. El uso de técnicas de procesamiento de textos científicos, datos disponibles en la Web y sitios personales bajo estos dominios sería de gran utilidad. Por lo tanto, el desarrollo de sistemas que realice el descubrimiento de conocimiento y representación de datos académicos-científicos a partir de textos disponibles es altamente benéfico para los usuarios que intentan localizar información de este tipo sobre grandes volúmenes de datos. En este artículo, se presenta un enfoque para descubrir relaciones semánticas entre los datos académicos y publicaciones científicas de un investigador. Además de representar la información descubierta en un modelo ontológico, el cual será de gran utilidad para la búsqueda de información precisa sobre datos académicos o de publicaciones científicas.

Palabras clave: Enlazando datos abiertos, publicaciones científicas, aprendizaje de relaciones, ontologías.

1 Introducción

Los investigadores producen publicaciones y tienen información adjunta, tanto académica como profesional, la cual expresa sus trayectoria en el pasado. Los investigadores utilizan la Web como medio para dar a conocer sus publicaciones y sus áreas de interés en la investigación. El conjunto de registros sobre publicaciones y perfiles en la Web constituye información valiosa que resulta de interés para los propios investigadores y para el público en general. Actualmente, esta información se encuentra disponible en diversos sitios o repositorios, tanto públicos como privados. Existen varios sitios o repositorios de información académica y sobre publicaciones, tales como: a) el sitio *Google Académico*[1] soportado por la empresa Google que contiene información sobre citas y perfiles especializadas en literatura científica-académica; b) *DBLP* [2] es un sitio web, patrocinado por la Universidad de Trier, que ofrece un

servicio para consultar información bibliográfica sobre revistas y memorias sobre el área de las ciencias computacionales; c) *CiteSeerX* [3] es una librería digital para publicaciones científicas y académicas, principalmente en los campos de la computación e informática, el cual incluye citas, documentos y estadísticas; d) *ArnetMiner* [4] es una plataforma que contiene publicaciones académicas, las cuales recolecta a partir de la Web, la cual permite encontrar conexiones entre investigadores.

Es posible acceder a la información sobre publicaciones científicas e información académica a través de la Web, mediante el uso de buscadores especializados, analizadores de páginas Web o bases de datos. Para que los usuarios localicen información oportuna y precisa sobre perfiles académicos o publicaciones científicas es necesario invertir mucho tiempo y esfuerzo, esto requiere un trabajo exhaustivo de análisis de resultados proporcionados por un buscador o el análisis directo de sitios Web. Aunado a esto, se tiene que los buscadores son propenso a proporcionar errores en sus resultados debido a la carencia de estructuras semánticas en los datos que ayuden a mejorar los resultados de las búsquedas.

Este trabajo se centra en resolver el problema de carencia de semántica con la finalidad de mejorar las estructuras de representación de los datos de publicaciones científicos y académicos de tal forma que se faciliten la localización de información oportuna. Las principales aportaciones de este trabajo son (a) la extracción de datos sobre publicaciones científicas a partir de recursos Web (sitios y páginas web) necesarios para descubrir nuevo conocimiento sobre ellos; (b) un enfoque para relacionar los datos sobre publicaciones científicas utilizando un modelo ontológico; y (c) la representación del conocimiento descubierto y sobre información académica de un investigador.

El resto del artículo está organizado de la siguiente manera. En la Sección 2 se presentan los trabajos relacionados al descubrimiento y representación de conocimiento y al procesamiento automático de textos científicos y sobre información académica. La Sección 3 presenta el modelo ontológico global. El enfoque propuesto en este artículo para el descubrimiento de conocimiento (relaciones entre los datos) se muestra en la Sección 4. El modelo ontológico se explica en la Sección 5, en la cual, además, se presenta la representación del conocimiento adquirido a partir de los textos utilizando un modelo ontológico, esta representación se apoya del poblado automático de la ontología sobre publicaciones científicas y datos académicos. Finalmente, las conclusiones y el trabajo a futuro son presentados en la Sección 6.

2 Trabajos relacionados

El descubrimiento de conocimiento es una tarea que ha aportado avances significativos en la representación (semi) automática de conocimiento y sobre todo en las búsquedas basadas en el significado (semántica) de la consulta. Por ello, se presentan trabajos que han hecho aportaciones en esta área. De esta manera, en [5] se presenta un enfoque donde aplican técnicas de minería de textos sobre información de publicaciones científicas y citas publicadas en *CiteSeer*^{X} con la finalidad de detectar investigadores influyentes en el área de las ciencias computacionales. [6] presenta a *DBconnect* una herramienta para la extracción de información a partir de grandes coleccio-

nes de datos estructurados, semiestructurados o no estructurados, específicamente, explota la base de datos de DBLP codificada en XML haciendo uso de un enfoque para extraer el conocimiento relevante acerca de la comunidad de investigación e incluso recomendar colaboraciones.

La herramienta *ArnetMiner*, reportada en [7] y [8], ha sido creada para la extracción y minería de información científica a partir de recursos disponibles en la Web. Esta herramienta aplica un análisis y minería de datos exhaustiva para investigadores a partir de recursos Web compartidos. Además, integra los datos de publicaciones en una red de bibliotecas digitales existentes modelando la red académica completa y finalmente, ofrece servicios de búsqueda sobre la red de datos de publicaciones integradas.

El trabajo presentado en [9] expone una herramienta que integra tres recursos de información, siendo capaz de agregarlos y exportarlos dentro de sus modelos ontológicos llamados *VIVO* y *CERIF*. Los autores presentan un caso de estudio en agricultura usando una base de datos bibliográfica enlazada a recursos Web con más de 7 millones de registros. Además, el método presentado combina datos provenientes de *Google Académico* para la información científica y agrega nueva información.

Además de las propuestas de extracción de conocimiento, se cuentan con diversos trabajos para la representación semántica de información sobre publicaciones científicas mediante el uso de ontologías. En este contexto, VIVO, expuesta en [10], es un modelo ontológico que incluye representación de personas, organizaciones y actividades involucradas en la investigación científica. Este modelo ontológico extiende ontologías existentes como *FOAF (Friend-of-a-Friend)*, la cual provee las bases para describir personas y organización, y la ontología bibliográfica (*BIBO*).

Otro modelo ontológico que se utiliza para la representación de información académica es BIBO [11] que ofrece un mecanismo para la representación semántica de datos bibliográficos en RDF y OWL. Más precisamente, se enfocan en presentar resultados eficientes sobre búsquedas científicas.

3 Modelo ontológico global

Esta sección presenta el esquema global ontológico que es utilizado para la representación de información sobre publicaciones científicas e información académica de un investigador.

Se utiliza la sintaxis de Manchester para OWL 1.1 [12] con el propósito de presentar el modelo ontológico global para la representación semántica de conocimiento en una sintaxis amigable para el usuario. Así pues, en modelo ontológico contiene las siguientes clases y subclases.

```
Class: Investigador
Class: Institución
Class: Articulo
Class: Publicacion
Class: PublicacionRevista
  SubClassOf: Publicación
```

```
  Class: MemoriaCongreso
    SubClassOf: Publicación
```

Estas clases se utilizan para representar instancias de publicaciones y su información relacionada con el autor y lugar donde se publica el artículo, el cual puede ser un congreso o en una revista. Además, se representan los autores y su información de lugar donde está adscrito como investigador.

El modelo ontológico incluye las siguientes propiedades de objeto, relaciones ontológicas que representan un vínculo entre dos clases.

```
ObjectProperty: fuePublicadoEn
    Domain: Artículo
    Range: Publicación
ObjectProperty: estaAdscritoA
    Domain: Investigador
    Range: Institución
ObjectProperty: fuePublicadoPor
    Domain: Artículo
    Range: Investigador
```

Finalmente, el modelo también incluye propiedades de tipos datos, las cuales se utilizan para agregar información de propiedades a las instancias, las cuales se presentan a continuación:

```
DataProperty: tieneTítulo
    Domain: Artículo
    Range: xsd:string
DataProperty: tienePagFinal
    Domain: Artículo
    Range: xsd:int
DataProperty: tieneNombreInstitucion
    Domain: Institución
    Range: xsd:string
DataProperty: tieneNombreInvestigador
    Domain: Investigador
    Range: xsd:string
DataProperty: tienePagInicial
    Domain: Artículo
    Range: xsd:int
DataProperty: tienePosiciónAcadémica
Domain: Investigador
    Range: xsd:string
DataProperty: tieneAnioDePublicación
   Domain: Articulo
```

```
   Range: xsd:int
DataProperty: tieneAreaInvestigación
    Domain: Investigador
    Range: xsd:string
DataProperty: tieneNombrePublicación
   Domain:Publicación
   Range: xsd:string
DataProperty: tieneCorreo
   Domain: Investigador
   Range: xsd:string
```

4 Descubriendo relaciones entre los datos

En esta sección se presenta el enfoque utilizado para el descubrimiento de relaciones semánticas entre los datos sobre perfiles científicos y académicos. En proceso completo de descubrimiento de relaciones semánticas incluye las siguientes tareas: localización y extracción de textos a partir de repositorio público de datos académicos; segmentación de los textos y etiquetado morfológico; y finalmente, la aplicación de patrones sintácticos (estructurales) para la localización de elementos (conceptos) académicos y relaciones entre ellos. En la Fig. 1, se puede observar el, no sólo el proceso de descubrimiento de relaciones entre los datos académicos, sino también la forma de representación de los mismos.

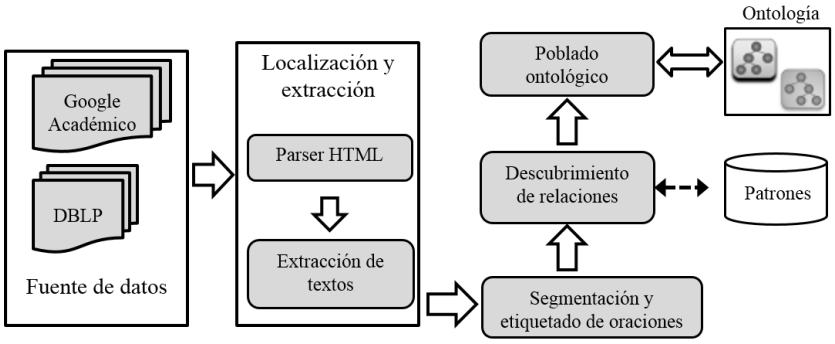

Fig. 1. Proceso de descubrimiento y representación de información académica.

4.1 Localización y extracción de textos

Los repositorios públicos sobre información académica son una fuente de datos valiosa. Por ello, se utiliza el repositorio de información académica *Google Académico* y *DBLP* para la localización de datos sobre un autor.

La idea es obtener los textos de citas e información académica de los repositorios mencionados con la finalidad de descubrir instancias de conceptos del dominio académico con sus propiedades y relaciones.

De esta manera, se procesa la página Web del investigador, tanto en el sitio *Google Académico* como en el sitio *DBLP* con la finalidad de extraer su información académica e información sobre sus publicaciones científicas. La Fig. 2 muestra un fragmento de texto extraído de Google Académico, el cual contienen información académica de un investigador, tal como: nombre, posición, áreas de investigación y correo electrónico.

Fig. 2. Fragmento de texto de *Google Académico*

Además, se utiliza la información sobre publicaciones científicas del sitio *DBLP*. En la Fig. 3 se despliega un fragmento de texto extraído del sitio *DBLP*, a partir del cual se puede identificar información relacionada a las publicaciones de un investigador.

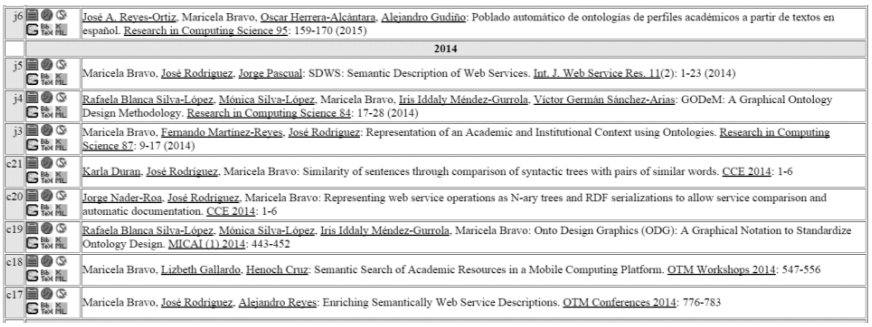

Fig. 3. Fragmento de texto del sitio *DBLP*

4.2 Segmentación y etiquetado morfológico

El texto extraído de *Google Académico* y *DBLP* es segmentado en oraciones con la finalidad de aplicar patrones sintácticos. Adicionalmente, un etiquetado de partes de oración es aplicado utilizando *TreeTagger* [13] bajo la arquitectura genérica para el procesamiento de texto llamada *GATE* [14].

El etiquetado de partes de oración, también llamado *POS tagger*, se encarga de asignar una categoría gramatical a cada palabra de una oración. En la Fig. 4 se muestra un ejemplo del etiquetado de una oración.

word	pos	lemma
The	DT	the
TreeTagger	NP	TreeTagger
is	VBZ	be
easy	JJ	easy
to	TO	to
use	VB	use
.	SENT	.

Fig. 4. Etiquetado de una oración utilizando *TreeTagger*

Esta información gramatical de las palabras es utilizada por los patrones para la identificación y extracción de relaciones semánticas entre los datos de las publicaciones científicas.

4.3 Descubrimiento de relaciones utilizando patrones

La información extraída se puede clasificar en dos (a) información académica: nombre del investigador, correo, posición y universidad de adscripción; (b) información científica, tal como: líneas de investigación o áreas de interés en investigación e información relacionada con las publicaciones de un investigador.

Con respecto a la información académica, se aplican patrones estructurales basados en etiquetas HTML para identificar datos académicos y crear relaciones con el investigador. La Tabla 1 muestra algunos patrones basados en etiquetas HTML para la extracción de información académica.

Tabla 1. Patrones para la identificación de información académica.

Relación identificada	Patrón
tieneCorreo	**Correo electrónico de verificación** <*input type ="text" name="u-v" value*=**Correo**>
tieneAreaInvestigación	**Áreas de interés** <*input type="text" name="u-i" value*= **AreaInv1 (, AreaInvN)*** >
tienePosiciónAcadémica	**Afiliación** * <*input type="text" name="u-a" value*= **PosiciónAcadémica** >

Por su parte la información científica, es extraída de DBLP y para la cual se aplican patrones semánticos basados en la información gramatical. De esta manera, se muestran un patrón para la identificación de información científica, las cuales se pueden observar en la Tabla 2.

Tabla 2. Patrones para la identificación de información científica.

Patrón	Patrón
1	NP (, NP)* : TITULO . PUBLICACIÓN (VOL)? : PagIni-PagFin (AñoPublicación)

Con el patrón mostrado en la Tabla 2 se extraen las relaciones de *tieneTítulo*, *tienePagFinal*, *tienePagInicial*, *fuePublicadoEn* y *fuePublicadoPor*. Es a partir de esta información (instancias y relaciones entre ellas) descubierta que se realiza el proceso de poblado automático del modelo ontológico.

5 Poblado automático de ontologías de perfiles científicos y académicos

El poblado automático de modelo ontológico se realiza a partir de las instancias de las clases descubiertas y las relaciones entre ellas. La idea principal del poblado automático crear instancias y relacionarlas en una ontología existente [15]. Bajo este contexto, en este artículo, el poblado automático de las ontologías de perfiles consiste en agregar instancias de clases y relaciones entre ellas dentro del modelo ontológico global existente.

Por lo tanto, esta fase toma las relaciones descubiertas y crea instancias ontológicas dentro de la clase correspondiente. Así pues, se crean instancias en la clase *Publicación*, *Investigadores*, *Artículo* y *Publicación* (*Revista* o *Congreso*). El siguiente código representa la creación de una instancia ontológica de un investigador y su información académica relacionada.

```
Individual:invReyes
    Types: Investigador
    Facts: estaAdscritoA instUAM
           tieneNombreInvestigador  "José  A.  Reyes-
    Ortiz"^^xsd:string
           tieneAreaInvestigación  "Procesamiento  de
    Lenguaje Natural"^^xsd:string
           tieneAreaInvestigación "Ingeniería Ontoló-
    gica"^^xsd:string
           tienePosiciónAcadémica "Doctor en Ciencias
    Computacionales"^^xsd:string
           tieneCorreo        "jaro@correo.azc.uam.mx"
    ^^xsd:string

Individual: instUAM
    Types: Institución
    Facts:  tieneNombreInstitución  "Universidad  Autó-
    noma Metropolitana"^^xsd:string
```

Además se crean instancias y relaciones entre ellas para la información científica del investigador. Como ejemplo se muestra el código generado en la sintaxis de Manchester para OWL 1.1, para la creación de instancias y relaciones de un investigador y su información científica, es decir, información de publicaciones.

```
Individual: artPoblado
    Types: Artículo
    Facts: fuePublicadoPor invReyes
           fuePublicadoEn revRSC95
           tieneTítulo "Poblado automático de ontolo-
    gías de perfiles académicos a partir de textos en
    español"^^xsd:string,
           tieneAnioDePublicación "2015"^^xsd:int,
           tienePagInicial "159"^^xsd:int,
           tienePagFinal "170"^^xsd:int

Individual: revRSC95
    Types: Revista
    Facts: tieneNombrePublicación "Research in Compu-
    ting Science 95"^^xsd:string
```

6 Conclusiones y trabajos futuros

Este artículo ha presentado un enfoque para la extracción y representación de conocimiento a partir de recursos Web compartidos. Las principales aportaciones de este artículo son (a) la extracción de datos sobre publicaciones científicas a partir de recursos Web (sitios y páginas web) necesarios para descubrir nuevo conocimiento sobre ellos; (b) un enfoque para relacionar los datos sobre publicaciones científicas utilizando un modelo ontológico; y (c) la representación del conocimiento extraído sobre información académica de un investigador.

A manera de resultados, se ha logrado crear instancias, de manera automática, para siete investigadores de la Universidad Autónoma Metropolitana, unidad Azcapotzalco. Para estos investigadores se ha identificado su información académica: correo, nombre de posición (puesto) y líneas de investigación. Un total de 85 instancias de artículos de investigación relacionadas a estos investigadores se han identificado y extraído de manera automática. Por cada instancia de artículo se ha extraído su información de quién lo pública, dónde se ha publicado (revista o congreso), año de publicación, página inicial y página final.

La información extraída y representada en el modelo ontológico puede ser de gran utilidad para apoyar diversas aplicaciones como sistemas de pregunta-respuesta o buscadores semánticos.

Como trabajo futuro, se planea la consideración de diversas fuentes de datos académicas como *CiteSeerX* o *ArnetMiner* para complementar y validar las publicaciones de científicas de los investigadores. Además, será necesario una validación de artículos duplicados (citados en dos sitios).

Referencias

1. Google Académico, https://scholar.google.com.mx/
2. The DBLP Computer Science Bibliography, http://dblp.uni-trier.de/
3. Scientific literature digital library and search engine: CiteSeer^{x}, http://citeseerx.ist.psu.edu/index
4. AMiner - Open Science Platform, https://aminer.org/
5. FIALA, D.: Extracting information from CiteSeer's textual data. Journal of Theoretical and Applied Information Technology 56 (2), 176-182 (2005)
6. Zaiane, O. R., Chen, J., Goebel, R.: DBconnect: mining research community on DBLP data. In: Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis, pp. 74-81. ACM, New York (2007)
7. Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., Su, Z.: Arnetminer: extraction and mining of academic social networks. In: Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 990-998. ACM, New York (2008)
8. Tang, J., Zhang, J., Zhang, D., Yao, L., Zhu, C., Li, J. Z.: ArnetMiner: An Expertise Oriented Search System for Web Community. In: Semantic Web Challenge. (2007)
9. Nogales, A., Sicilia, M. A., Jörg, B.: Combining VIVO and Google Scholar data as sources for CERIF Linked Data: a case in the agricultural domain. In: Procedia Computer Science, vol. 33, pp. 266-271. (2014)
10. Lezcano, L., Jörg, B., Sicilia, M. A.: Modeling the context of scientific information: Mapping VIVO and CERIF. In: Advanced Information Systems Engineering Workshops, pp. 123-129. Springer, Heidelberg (2012)
11. Dimić Surla, B., Segedinac, M., & Ivanović, D. A.: BIBO ontology extension for evaluation of scientific research results. In: Proceedings of the Fifth Balkan Conference in Informatics, pp. 275-278. ACM, New York (2012)
12. Horridge, M., Patel-Schneider, P. F.: Manchester syntax for OWL 1.1. OWL: Experiences and Directions, Washington (2008)
13. Helmut S.: Improvements in Part-of-Speech Tagging with an Application to German. In: Proceedings of the ACL SIGDAT-Workshop, pp. 47-50. Dublin (1995)
14. Cunningham, H.: GATE, a general architecture for text engineering. Computers and the Humanities 36 (2), 223-254 (2002)
15. Buitelaar, P., Cimiano, P.: Ontology learning and population: bridging the gap between text and knowledge – Volume 167 Frontiers in Artificial Intelligence and Applications. IOS Press, Amsterdam (2008)

Capítulo 2

Análisis de Sentimientos Basado en Aspectos

Orlando Ramos, David Pinto, Darnes Vilariño, Mireya Tovar

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación
Blvd. 14 Sur y Av. San Claudio, Col. San Manuel, Puebla, Pue., México.

orlandxrf@gmail.com, {dpinto,darnes,mtovar}@cs.buap.mx

Abstract. Con el paso de los años el internet ha ido avanzando de una forma vertiginosa y con ello se han ido creando diversas plataformas como son: redes sociales, blogs de diferentes temas, plataformas de aprendizaje, plataformas informativas, wikis, plataformas para dependencias gubernamentales, universidades, etc. en donde acceden los usuarios para hacer uso de estas, e interactúan con otros usuarios ya sea dentro de la misma plataforma, fuera de ella o entre múltiples plataformas, generando una gran cantidad de información que ya es procesada por diferentes plataformas web para detectar usuarios potenciales para los servicios o productos que ofrecen estas empresas. Presentamos un estudio sobre los trabajos realizados por investigadores del Procesamiento del Lenguaje Natural en el área del análisis de sentimientos, minería de opinión y más concretamente en el análisis de sentimientos basado en aspectos, donde se abordan las técnicas y procedimientos usados por cada autor que se han desarrollado en los últimos años.

Keywords: analisis, sentimientos, minería, opinión, aspectos, polaridad, positiva, negativa, neutra, NRE, OTE, bigramas.

1 Introducción

El análisis de sentimientos o minería de opinión se ha convertido en una de las principales herramientas, para que empresas obtengan con la ayuda del Procesamiento de Lenguaje Natural (PLN), los recursos necesarios de sus clientes (usuarios) los cuales dejan comentarios, recomendaciones y opiniones en sus sitios web, blogs, en las redes sociales, e incluso en algún otro tipo de medio electrónico en el cual se recaben información de los usuarios, para así de este modo explotar está información en beneficio propio.

Es una práctica común que las compañías que venden productos en la Web pidan a sus clientes que revisen los productos y servicios asociados. Como el comercio electrónico se está volviendo más y más popular, el número de críticas u opiniones de clientes que recibe un producto crece rápidamente. Para un producto popular, el número de comentarios puede estar en cientos. Esto hace que sea difícil de leerlos para un cliente potencial con el fin de tomar una decisión sobre si comprar el producto o no[3].

Entre una de las múltiples tareas de las que se ocupa el PLN se encuentra la clasificación de textos, que consiste en la asignación de un conjunto de categorías a una colección de documentos, resolviéndose de esta forma la clasificación objetiva de documentos. Existe una gran cantidad de textos en donde el contenido subjetivo es lo más relevante, y cuyo procesamiento no debería limitarse a aplicar únicamente las técnicas de la clasificación de documentos. El análisis de sentimientos trata de clasificar los documentos en función de la polaridad de la opinión que expresa su autor. Esta nueva área que combina PLN y minería de textos, incluye una gran cantidad de tareas que han sido tratadas en mayor o menor medida[1].

Dentro del campo computacional el análisis de sentimientos puede referirse de diferentes formas, dependiendo del termino que se encuentre en boga, como son: opinión, sentimiento y subjetividad en textos. Aunque también se puede referir como minería de opinión, análisis de sentimientos y/o análisis de la subjetividad. Otras veces se le llama también revisión de sentidos, extracción de valoración y computación afectiva[5].

2 Análisis de Sentimientos

En esta sección abordaremos los trabajos relacionados con la minería de opinión y el análisis de sentimientos basados en aspectos.

2.1 Análisis usando polaridad y minería de opinión

En el trabajo de [4] recopilaron un corpus de 300,000 tweets los cuales están clasificados en tres polaridades: los textos que contienen las menciones positivas, como la felicidad, la diversión o la alegría, los textos que contienen emociones negativas, tales como la tristeza, la ira o la decepción y los textos objetivos que sólo declaran un hecho o no expresan una emoción. Para la recolección de tweets de emociones positivas se consultaron los emoticones felices ":-)", ":)", "=)", ":D", etc. y para los tweets con emociones negativas se basaron en la consulta de emoticones del tipo ":-(", ":(", "=(", ";(", etc. Estos dos tipos de corpora fueron utilizados para entrenar un clasificador para reconocer sentimientos positivos y negativos. En el caso de los tweets objetivos (no tienen emociones) se recolectaron de cuentas de Twitter de periódicos y revistas populares, tales como "New York Times", "Washington Posts", etc. consultando cuentas de 44 periódicos para recoger un conjunto de entrenamiento de textos objetivos. Primeramente etiquetaron los corpus con TreeTagger[10] para el inglés, después analizaron la objetividad (el conjunto neutro) y subjetividad (la mezcla de los aspectos positivos o negativos) de las etiquetas POS obtenidas donde se observó que los textos objetivos tienden a contener más los nombres comunes y nombres propios (NPS, NP, NNS), mientras que los autores de los textos subjetivos utilizan más a menudo los pronombres personales (PP, PP$). Para los experimentos usaron unigramas, bigramas, trigramas y POS en un clasificador Naïve Bayes. Para incrementar la precisión de la clasificación se eliminaron los n-gramas comunes (n-gramas que no indican

ningún sentimiento ni objetividad) a través de dos estrategias, la primera se basa en el cálculo de la entropía de la distribución de probabilidad de la aparición de un n-grama en diferentes conjuntos de datos (diferentes polaridades), donde se optó por los valores bajos de entropía que indican que un n-grama aparece en algunos de los conjuntos de datos de sentimientos con más frecuencia que otros y por consiguiente lograron destacar un sentimiento en particular (o de objetividad). La segunda estrategia consiste en introducir un término "salience" (rasgo sobresaliente) que se calcula para cada n-grama, la medida introducida toma un valor entre 0 y 1, donde los valores bajos indican una baja relevancia del n-grama, por lo que el n-grama se descrimina y solo se toman los valores altos, caso contrario al de la entropía. Consiguiendo los mejores resultados al utilizar bigramas, los cuales proporcionan un buen equilibrio entre una cobertura de unigramas y una capacidad de capturar los patrones de expresión de sentimientos en trigramas.

El resumen de opinión basado en características de opiniones de usuarios sobre productos vendidos en línea, es donde [3] centra su investigación, primeramente se identifican las características del producto en los cuales los clientes han expresado opiniones y se clasifican las características de acuerdo a la frecuencia en la que aparecen en las críticas. Posteriormente para cada característica se identifican el número de críticas de los clientes que tienen opiniones positivas o negativas. Antes de aplicar las técnicas, se descargan las críticas en un crawler y/o base de datos, para después etiquetarlas (POS) con la ayuda del NLProcessor Linguistic Parser enfocándose en encontrar características que aparecen explícitamente como sustantivos o frases nominales. Además el pre-procesamiento incluye la eliminación de stop words, truncamiento (steemming) y coincidencias aproximadas (fuzzy matching). Después se realiza una generación de características frecuentes es decir, se buscan las características en las que la gente esta más interesada, con reglas de minería de asociación donde se tiene un conjunto de elementos y un conjunto de transacciones (cada transacción consiste de un subconjunto de elementos). Se continua con una poda de características, pues no todas las características frecuentes generadas por la minería de asociación son útiles o auténticas, además se podan las redundantes o que no son interesantes. Continuando así con la extracción de palabras de opinión que son las palabras que la gente usa para expresar una opinión positiva o negativa, la forma en la que se realiza la extracción es tomando una característica del producto dentro de una oración, donde se sabe que las palabras que rodean está característica son palabras de opinión, de este modo se pueden extraer palabras de opinión de la base de datos utilizando todas las características frecuentes restantes (después de la poda). Una técnica aplicada más es la identificación de características poco frecuentes es decir, las características que sólo un puñado de personas expresan, en cuanto a su extracción para cada oración si esta no contiene características frecuentes, pero si contiene una o más palabras de opinión se busca el sustantivo/frase nominal más cercanas a ellas y se almacena en el conjunto de características poco frecuentes. Una vez que se han aplicado estás técnicas y se han identificado las características de opinión se determina la orientación semán-

tica usando una técnica de muestreo y WordNet[2] y de está forma deciden la orientación de opinión para cada oración. Los mejores resultados se alcanzaron con la técnica que identifica el número de veces que aparece la característica del producto en un sustantivo o frase nominal.

2.2 Análisis de Sentimientos basados en aspectos

Uno de los trabajos recientes sobre este tópico es presentado por [6] en la Task12 del SemEval 2015. En donde a partir del trabajo [7], dada una reseña por un usuario sobre una entidad *e,* el objetivo fue identificar todos los aspectos (términos explícitos o categorías) y las polaridades correspondientes. Un aspecto (término o categoría) puede ser indicado como: una parte/componente de *e*, un atributo de *e*, o un atributo de una parte/componente de *e*.

Por lo que en [6] se crea un framework para el ABSA (Aspect Based Sentiment Analysis) en donde se plantea una categoría de aspecto definida como una combinación de una entidad de tipo E y un atributo de tipo A. Donde E puede ser la reseña de una entidad *e* en sí misma (ejemplo: laptop), una parte/componente de *e* (ejemplo: batería o atención al cliente), u otra entidad relevante (ejemplo: el frabricante de *e*), mientras que A es un atributo particular (ejemplo: calidad, durabilidad) de E. E y A son nombres de conceptos (clases) de una ontología de un determinado dominio y no necesariamente ocurren como términos en una oración.

Para esta tarea se contemplan dos subtareas, la primera es en el dominio del ABSA en el cual dado un texto de reseñas sobre una laptop o restaurante, se identifican todas las tuplas de opinión con los siguientes tipos de información: *categoría de aspectos* (identificar los pares: entidad E y atributo A, cuya opinión es expresada en un texto dado), *OTE: opinión de una expresión objetiva* (extraer la expresión lingüística usada en el texto dado para referirse a la entidad E, de cada par E#A) y *polaridad de sentimientos* (para cada par E#A identificado se le asigna una etiqueta: positiva, negativa, neutra). Para la segunda subtarea que se encuentra fuera del dominio del ABSA los participantes tuvieron la oportunidad de probar sus sistemas en el dominio de reseñas de hoteles.

El conjunto de datos fue previsto para tres dominios (laptops, restaurantes y hoteles) en un formato XML. Cada conjunto de datos fue anotado por un lingüista usando BRAT, una herramienta de anotaciones basada en web que fue configurada apropiadamente para las necesidades de esta tarea. Para las evaluaciones en la fase A se usó la medida F-1 para categoría de aspecto y extracción de opiniones objetivas y para la fase B se evaluó la polaridad de sentimientos.

Para la extracción de *categorías de aspectos* (E#A) se entrenó una máquina de soporte vectorial (SVM) con un núcleo lineal, en particular n unigramas de características son extraídas de la respectiva oración de cada tupla que es encontrada en el conjunto de entrenamiento. El valor de la categoría de la tupla es usado como la etiqueta correcta del vector de características. De forma similar para cada conjunto de prueba de oraciones s, un vector de características es construido y el entrenamiento de la SVM es usado para predecir las probabilidades

de asignación de cada posible categoría a *s*. De este modo un umbral t es usado para decidir cual de las categorías sera asignada a *s*.

En el caso del entrenamiento de reseñas para cada *categoría c* se creó una lista de *opiniones de expresiones objetivas OTE*. Estas son extraídas de (el entrenamiento) las tuplas de opinión cuyo valor de la categoría es *c*. Dada una oración de prueba *s* y una categoría asignada *c* se busca en *s* la primera ocurrencia de cada OTE de *c* en la lista, a OTE se le asigna la primera ocurrencia encontrada en *s,* si no se encuentran ocurrencias se coloca NULL.

Para la predicción de *polaridad* se entrenó un clasificador SVM con un núcleo lineal, al igual que en la extracción de *categorías de aspectos n* unigramas de características son extraídas de la respectiva oración de cada tupla que es encontrada en el conjunto de entrenamiento. Adicionalmente a cada par de categoría (E#A) se le asigna un valor entero distinto que indica que la categoría de la tupla es usada. La etiqueta correcta para el vector de características extraído del entrenamiento es el valor correspondiente de polaridad. Por lo que para cada tupla {categoría, OTE} de una oración de prueba *s*, un vector de características es construido y es clasificado usando la SVM.

En cuanto a los resultados de la fase A, el mejor equipo para la extracción de *categorías de aspectos,* modeló el problema como una clasificación multiclase con características basadas en n-gramas, análisis sintáctico y de clusters de palabras aprendidas de Amazon y Yelp (de laptos y restaurantes). El segundo mejor equipo utilizó un clasificador MaxEnt independiente con bolsa de palabras de características (palabra, lema, etc.) para cada entidad y para cada atributo. En cuanto a la extracción de *opiniones de expresiones objetivas OTE* el mejor equipo abordó el problema usando un averaged perceptron con un esquema de etiquetado BIO, las características incluían n-gramas, clases de tokens, prefijos y sufijos de n-gramas y cluster de palabras aprendidas de datos adicionales Yelp (clusters Brown and Clark), Wikipedia (clusters word2vec). El segundo mejor equipo se basó en un modelo de campos aleatorios condicionales (CRF) con características basadas en cadenas de palabras (obtenidas a partir de árboles sintácticos) listas de nombres (extraídas por medio de la frecuencia) y clusters de Brown.

Para los resultados de la fase B sobre la *polaridad de sentimientos* el mejor equipo utilizó un clasificador MaxEnt junto con características basadas en n-gramas, etiquetado POS, lematización, negación de palabras y un lexicón de sentimientos.

En el trabajo de [11] para resolver la Task12 del SemEval 2015 (ASBA) usan en su sistema dos algoritmos de aprendizaje automático supervisado: una red de prealimentación sigmoidal para entrenar clasificadores binarios para la clasificación de categorías de aspectos y un algoritmo de campos aleatorios condicionales para entrenar clasificadores para la extracción de opiniones objetivas. Para el preprocesamiento de los datos se tokenizaron y se analizaron con Stanford Parser. En cuanto a la identificación de características las dividieron en cinco subcategorias para su procesamiento quedando de la siguiente manera: *Word*, cada palabra es usada como una característica, para la extracción de

opiniones objetivas la palabra anterior y la posterior fueron usadas como características. *Bigram*, todos los bigramas de palabras encontrados en una oración son usados como características. *Name List*, para el dominio de restaurantes, extrajeron dos listas de nombres de alta precisión del conjunto de datos de entrenamiento y se usaron para las pruebas de adhesión. En la primera lista se extrajeron solamente las opiniones objetivas con alta frecuencia. En la segunda lista se tomaron en cuenta los conteos de palabras individuales en las opiniones objetivas y se guardaron las palabras que ocurren frecuentemente en el conjunto de entrenamiento como parte de una opinión objetiva. *Head word*, de las oraciones del árbol de análisis sintáctico se extrae la palabra principal de cada oración y es usada como característica. *Word Cluster*, se introducieron clusters de Brown y clusters K-means de dos diferentes fuentes de datos sin etiquetar: Multi-Domain Sentiment Dataset y Yelp Phoenix Academic Dataset. Para el dominio de los restaurantes se generaron listas de nombres de posibles opiniones objetivas usando el algoritmo de doble propagación[8], dado que las reglas sólo pueden identificar una sola palabra se modificaron las reglas para poder incluir palabras nominales consecutivas antes de la palabra buscada.

En sus experimentos para la clasificación de categorías de aspectos se basan en un conjunto clasificadores binarios uno-contra-todos. Un clasificador para cada categoría encontrada en el conjunto de entrenamiento, para cada oración en el conjunto de entrenamiento se extrajeron las características de todas las palabras en la oración para crear ejemplos de entrenamiento. Donde la etiqueta del ejemplo depende sobre que categoría C se está entreanando: (1) si la oración contiene a C como una de estas categorías y (-1) en otro caso. Para la extracción de categorías objetivas se modelo como una tarea de clasificación secuencial, donde a cada palabra de la oración se le asigna una etiqueta utilizando el esquema IOB2.

Los mejores resultados que obtuvieron en la competencia con el sistema con restricciones son *Word*, *Bigram* y *Name List,* aunque los sistemas propuestos sin restricciones alcanzaron mejores rendimientos que los sistemas con restricciones para todas las evaluaciones, lo que indica que el uso de recursos externos beneficia el rendimiento.

El sistema que presentó [9] en el SemEval 2015 fue el que alcanzó la precisión más alta en la extracción de polaridad de sentimientos en el dominio de laptops y restaurantes. En su sistema ellos utilizaron un clasificador de aprendizaje automático supervisado combinado con un proceso de selección basado en la probabilidad, para entidades y la detección de atributos en la categoría. Y para la detección de categorías objetivas se realizó un catálogo de entidades que se llenó con objetivos conocidos para cada tipo de entidad, llamadas NER (reconocimiento de entidades nombradas). Y para lo polaridad de sentimientos se utilizó un clasificador de aprendizaje automático supervisado, teniendo una bolsa de palabras (BoW), lemas, bigramas después de verbos, puntuación basada en características, junto con un lexicón basado en características.

Para realizar sus experimentos, primeramente etiquetaron las reseñas de los usuarios asignándoles 0 o más tipos de entidades, además de etiquetar sus atrib-

utos, para después elegir y combinar las entidades y atributos identificados para formar un aspecto (aspect annotation). Cuando analizaron los datos de entrenamiento, encontraron que en algunas oraciones puede haber opiniones sobre diversos tipos de entidades o atributos. Por lo cual se optó para entrenar un clasificador para cada tipo de entidad, y un clasificador para cada etiqueta de atributo. El proceso de entrenamiento fue el mismo para todas las etiquetas, tipos de entidad o etiquetas de atributo, de cada dominio. El siguiente paso fue crear un conjunto de datos donde cada instancia es una oración del texto y su clase es *tag* (si la sentencia tenía por lo menos una opinión con esa etiqueta) de lo contrario *no_tag*. Cuando una oración tiene *no_tag* el sistema asume que no es una opinión, en el caso de que contenga 1 etiqueta en el tipo de entidad y 1 etiqueta en la etiqueta del atributo, entonces se dice que es un caso trivial, donde la unión de los dos resultados están dentro del aspecto (aspect annotation). Para las oraciones con 1 etiqueta en el tipo de entidad y 0 etiquetas en la etiqueta del atributo, el sistema busca el aspecto (aspect annotation) más frecuente dentro del dominio de la oración, que incluye este tipo de entidad. De forma equivalente se realiza en el caso donde las oraciones tienen 0 etiquetas en el tipo de entidad y 1 etiqueta en la etiqueta del atributo. Pero si ambas tienen más de 1 etiqueta dentro (tipo de entidad y atributo) el sistema aplica un ciclo, donde en cada iteración forma el par más frecuente (entidad, atributo) en ese dominio y elimina estas dos etiquetas de la oración.

Se recolectaron OTE de cada tipo de entidad de los datos de entrenamiento formando un catálogo. Si alguna opinión objetiva ya conocida aparece junto a un verbo o adjetivo, se elige como OTE. Si no se puede elegir un OTE el sistema aplica un reconocimiento de entidades nombradas, en busca de referencias de la organización y localización de entidades usando Stanford NER tool[1]. Cuando se encuentra algún OTE, con ayuda del catálogo o con la herramienta NER, se marca el texto y la posición, en caso de no encontrarse alguna OTE se coloca el valor de *NULL*.

En cuanto a la polaridad de sentimientos, se utilizó un clasificador de aprendizaje automático supervisado para predecir cada polaridad de opinión (positiva, negativa, neutra). En este caso sólo fue un clasificador, en el modelo las oraciones que no tienen una opinión no se consideran en el entrenamiento, ya que la polaridad está asociada con las opiniones. En el clasificador, para cada opinión se creó una instancia de polaridad, conteniendo la oración del texto, su dominio, su aspecto (aspect annotation) de tipo de categoría y atributo, OTE (si está disponible), y la polaridad de opinión. Debido a la utilización de lexicones de sentimientos este sistema funciona en la categoría sin restricciones de la Task12.

Los mejores resultados obtenidos por parte de [9] en SE-ABSA15 son en la detección y extracción de polaridad de sentimientos, alcanzando un 79.34 de precisión en el dominio de laptops y un 78.69 de precisión para el dominio de restaurantes, esto en la categoría sin restricciones.

[1] http://nlp.stanford.edu/software/CRF-NER.shtml

3 Conclusiones

El ABSA (Aspec Based Sentiment Analysis) es una de las líneas de investigación en la cual los investigadores proponen diferentes técnicas para su análisis, y que ha alcanzado niveles de rendimiento aceptables en la extracción de características de reseñas de usuarios. En las tareas de SemEval es donde se presentan los mayores avances, donde los investigadores han propuesto técnicas como son los algoritmos de aprendizaje automático, redes neuronales, máquinas de soporte vectorial, campos aleatorios condicionales (CRF), apoyándose de herramientas para el etiquetado, y para el análisis sintáctico.

Un punto clave y que ayuda en gran medida a incrementar la precisión y obtener puntuaciones más altas, es apoyarse de conjuntos de datos externos, es decir que no proporcionan en el SemEval, para realizar entrenamientos de los sistemas que serán propuestos y con esto tener un sistema que sea competitivo y obtenga los resultados esperados.

Dentro del conjunto de características que más usan los autores y que les ha dado buenos resultados están los bigramas, POS, lemas, OTE, NRE, y lexicones de datos, por mencionar algunos. Ya que estás características son las que permiten la detección y extracción de sentimientos en la mayoría de los artículos mencionados.

References

1. Eugenio Martınez Cámara, MaTeresa Martın Valdivia, and L Alfonso Urena. Análisis de sentimientos. *IV Jornadas TIMM Tratamiento de la Información Multilingüe y Multimodal*, page 61, 2011.
2. Christiane Fellbaum. *WordNet*. Wiley Online Library, 1998.
3. Minqing Hu and Bing Liu. Mining opinion features in customer reviews. In *AAAI*, volume 4, pages 755–760, 2004.
4. Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.
5. Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.
6. Maria Pontiki, Dimitris Galanis, Haris Papageorgiou, Suresh Manandhar, and Ion Androutsopoulos. Semeval-2015 task 12: Aspect based sentiment analysis. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 486–495, Denver, Colorado, June 2015. Association for Computational Linguistics.
7. Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. Semeval-2014 task 4: Aspect based sentiment analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 27–35, Dublin, Ireland, August 2014. Association for Computational Linguistics and Dublin City University.
8. Guang Qiu, Bing Liu, Jiajun Bu, and Chun Chen. Opinion word expansion and target extraction through double propagation. *Computational linguistics*, 37(1):9–27, 2011.
9. José Saias. Sentiue: Target and aspect based sentiment analysis in semeval-2015 task 12. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 767–771, Denver, Colorado, June 2015. Association for Computational Linguistics.
10. Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of the international conference on new methods in language processing*, volume 12, pages 44–49. Citeseer, 1994.
11. Zhiqiang Toh and Jian Su. Nlangp: Supervised machine learning system for aspect category classification and opinion target extraction. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 496–501, Denver, Colorado, June 2015. Association for Computational Linguistics.

Capítulo 3

Extracción automática de relaciones semánticas en corpus de dominio.

Hugo Chávez, Mireya Tovar

Benemérita Universidad Autónoma de Puebla, Blvd. 14 Sur y Av. San Claudio, Col. San Manuel, Ciudad Universitaria, Puebla, Pue., México.

hugo.lasserre@anahuacxalapa.mx,mtovar@cs.buap.mx

Abstract. En el presente documento se aborda un estudio acerca de los trabajos realizados por diferentes investigadores en el área de Procesamiento de Lenguaje Natural enfocado al análisis y extracción de relaciones semánticas en corpus de dominio. Tomando en cuenta métodos utilizados por la comunidad global en procesamiento de lenguaje natural como por ejemplo los métodos basados en diccionarios que parten de la idea de que una relación semántica de hiperonimia se puede encontrar en la primera frase del significado de una palabra en especifico. Métodos basados en agrupamiento que afirman que textos o palabras similares comparten contextos similares por lo cual la relación entre ellos está en el contexto que tenga el texto o palabra. Y métodos basados en patrones que son reglas ya definidas que se tienen que cumplir para encontrar ciertas relaciones.

Keywords: relaciones semánticas, extracción automática, taxonomías

1 Introducción

En los últimos años, ha surgido la necesidad de procesar y/o clasificar información de manera automática debido al crecimiento acelerado de la información disponible en Internet, empresas, organizaciones y repositorios en general. Este tipo de procesamiento requiere que la información sea representada de tal manera que sea entendible por las computadoras para que dicho procesamiento se pueda realizar de manera automática.

La importancia de las relaciones semánticas aplicadas a este trabajo radica en la necesidad de clasificar información de manera automática, es decir, necesitamos saber por ejemplo si un cierta información en un texto habla del mismo tema que otra parte de información en un texto totalmente diferente.

Hoy en día muchas aplicaciones de procesamiento de lenguaje natural hacen uso de tesauros, listas de palabras o de términos, empleados para representar conceptos como el sistema WordNet [1], que sirve como un diccionario de conocimiento léxico para el procesamiento de la semántica de palabras y documentos.

En el presente trabajo se abordan las diferentes maneras en que se encuentran relaciones semánticas de manera automática dentro de textos que pueden

estar estructurados o no estructurados. Específicamente, conceptos que están englobados dentro de otros conceptos. La siguiente sección considera los métodos basados en diccionarios (sección 3.1), métodos basados en agrupamiento (sección 3.2) y métodos basados en patrones (sección 3.3) siendo estos los mas estudiados dentro del área de extracción automáticas de relaciones semánticas.

2 Relaciones Taxonomicas

El procesamiento de lenguaje natural (PLN) consiste en la utilización de un lenguaje natural para comunicarnos con la computadora, un lenguaje natural es aquel que ha evolucionado con el tiempo para fines de comunicación humana, como el español. Una computadora debe entender las oraciones que le sean proporcionadas, el uso de estos lenguajes naturales facilita el desarrollo de programas que realicen tareas que sean basadas en el lenguaje o bien desarrollar modelos que ayuden a comprender los mecanismos humanos relacionados con el lenguaje. [2]

El PLN cuenta con una arquitectura de diferentes niveles como son el nivel de integración del discurso, nivel pragmático, semántico, sintáctico, morfológico, y fonológico. Nos enfocaremos en el nivel semántico el cual estudia la codificación del significado dentro de las expresiones lingüísticas. [3]

Dentro del nivel semántico encontramos relaciones conocidas como "es-un", en ingles "is-a", la cual es una relación entre clases donde una clase A es subclase de otra clase B, dentro de esta relacion encontramos otra, la relacion de hipónimo e hiperónimo las cuales son relaciones entre tipos de clases definiendo una relación taxonómica donde mediante una relación de herencia un hipónimo tiene una relación de "tipo-de" (es-un) con su hiperónimo. Por ejemplo, el hiperónimo de "lunes y martes" es "días de la semana" es decir, lunes es un tipo de día de la semana y martes también es un tipo de día de la semana. [4]

3 Extracción Automatica de relaciones semanticas

En esta sección se abordaran diferentes conceptos, tesis, algoritmos y trabajos enfocados a la extracción automática de relaciones semánticas dentro de textos y/o corpus de dominio, haciendo énfasis en las relaciones semánticas llamadas Taxonomías enfocadas a la hiponimia (inclusion semántica de un término en otro), las cuales son usadas amplia mente para organizar conocimiento de manera ontológica usando relaciones de generalización y especialización a través de una herencia simple o múltiple que puede ser aplicada para la clasificación de información.

Se espera que un sistema o método para la obtención automática de relaciones semánticas cumpla con las siguientes características:

- Rendimiento: Debe de generar relaciones de alta precisión.
- Supervisión Mínima: Debe de requerir mínima interacción humana o ninguna.
- Generalidad: Debe ser aplicable a diferentes tipos de relaciones.

De acuerdo a Ortega Mendoza [5], los trabajos que tratan la relación de hoponimia usan diversas técnicas para realizar la tarea. Entre los enfoques mas comunes se encuentran: los métodos basados en diccionarios, métodos basados en agrupamiento y métodos basados en patrones.

3.1 Métodos basados en diccionarios

Estos métodos asumen que los diccionarios ya están en un formato legible por una maquina y que contienen conocimiento explicito de manera estructurada. Parten de la idea de que el hiperónimo de una palabra puede aparecer en la primera frase nominal de la definición de la misma. Por ejemplo: Primavera: "La estación entre invierno y verano en la cual aparecen flores". Ahí podemos extraer "estación" como hiperónimo de "primavera" por estar en la primera frase nominal de la definición.

Este tipo de métodos son muy precisos pero presentan ciertas desventajas. Los diccionarios no contemplan términos específicos de un dominio como los corpus, casi siempre son términos muy generales y de diferentes dominios.

En el trabajo de [6] se usan datos disponibles en DBPedia.org para construir un conjunto de definiciones de términos en ingles. Para cada concepto que se obtiene del aritculo que esta siendo analizado en ese momento, un par (c,d) es construido donde C es el titulo exacto de un articulo en Wikipedia y d la definición del articulo. La extracción automática de relaciones semánticas basadas en diccionarios tiene su principio en que palabras similares tienen definiciones similares. El método propuesto usa una medida de similitud que toma como entrada un conjunto de conceptos y da como resultado relaciones entre ellos. Por ejemplo, el conjunto de términos (cocodrilo,animal,construcción,casa) daría como resultado (cocodrilo,animal),(construcción,casa) etc. Lo que quiere decir que el cocodrilo es un animal y una casa es una construcción, un concepto abarca al otro.

3.2 Métodos basados en agrupamiento

Esta técnica es la que suele dar mejores resultados, dado que para construir las relaciones se requieren ciertos datos de entrada que se pueden recoger en una practica de campo y así tener características o clasificaciones mas especificas y poder detectar de mejor manera cuando hay una relación.

Los métodos basados en agrupamiento toman como base la hipótesis de Harris citada por Cimiano (2006), la cual indica que las palabras similares comparten contextos similares. Gracias a este enfoque las palabras se caracterizan por su contexto y se agrupan de acuerdo con la similitud entre contextos.

Los autores de [7] han experimentado con un corpus basado en métodos para construir relaciones semánticas semiautomáticas. Su sistema usa un corpus de texto y un conjunto de palabras "semilla" para cada categoría y así poder identificar otras palabras que también pertenezcan a esa categoría. El algoritmo usa estadísticas simples para generar una lista ordenada por ranking de posibles palabras para cada categoría.

De acuerdo a [8] diferentes métodos han sido propuestos en la literatura para atacar el problema de obtener la derivación jerárquica de un texto de manera semiautomática o automática y estas pueden ser agrupadas en dos clases, los algoritmos basados en similitud y los conjuntos teóricos.

El primer tipo de método se caracteriza por el uso de una medida de similitud o distancia con el fin de calcular la similitud por parejas entre los vectores correspondientes a los términos, con el fin de decidir si son semánticamente similares y por lo tanto ser agrupados o no. Mas a fondo estos métodos pueden ser categorizados en métodos de aglomeración (bottom-up) y de division (top-down) que son estrategias de procesamiento de información.

El trabajo de [9] aborda la co-ocurrencia de términos en un corpus para la extracción de relaciones semánticas. Esta técnica asume que el hiperonimo de un término será encontrado en los términos que ocurran mas veces cerca del termino en el corpus. Su estrategia es la siguiente:

- Se toman términos como semilla de la taxonomía a ser construida. Un termino semilla es un termino que sirve como punto de construcción de una taxonomía, puede ser cualquier termino, que tenga el mayor peso para el dominio de la taxonomía o que se repita lo suficiente en los documentos a analizar.
- Se analizan las relaciones léxicas de los términos inspeccionando cuales son los términos con una mayor co-ocurrencia con los términos semilla. Se analiza el primer orden de co-ocurrencia, que busca la coocurrencia de términos dada una ventana de contexto.
- Después se busca el segundo grado de coocurrencia que se refiere a buscar la relación de entre un termino A con C cuando A coocurre con el término B y B también co-ocurre con C.
- En la ultima etapa los términos son ordenados en una taxonomía.

3.3 Métodos basados en patrones

A lo largo de los años se ha visto un considerable trabajo relacionado a la extracción de información basada en patrones. Hearts (1992) fué la pionera en el uso lexico-sintactico de patrones para la extraccion automatica de relaciones semanticas. Ella encontraba relaciones de hiponimia basadas en un pequeño conjunto de patrones previamente definidos como "X,...,Yy/o otra Z" y tambien patrones como "Z como X y/o como Y". [10]

El articulo [11] identifica un método para reconocer patrones lexico-sintacticos. Esto implica la búsqueda de términos específicos que están conectado mediante alguna relación semántica y derivando posibles patrones de los resultados en un corpus. Si estos patrones devuelven de manera correcta relaciones entonces estos pueden ser aplicados independientemente del dominio en el que se quiera aplicar para identificar y extraer definiciones. Patrones Lexico-Sintacticos pueden modelar diferentes relaciones pero hiponimia ha dado los mejores resultados desde 1992.

Estos métodos se apoyan en la idea de que existen frases, convenciones o estilos de palabras que las personas repiten al momento de relacionar un homónimo con si hiperónimo dentro de un texto. Estos patrones si ya se encuentran registrados pueden permitirnos extraer instancias de la relación de hiponimia al aplicarse a un corpus.

Las primeras pruebas bajo patrones que se realizaron fueron construidas manualmente, es decir, después de observar la forma en la que los conceptos se describen y relacionan en un texto, un experto de dominio identificaba y formaba un conjunto de patrones sintácticos para crear una pareja hipónimo-hiperónimo.

Estos métodos se apoyan en la idea de que existen frases, convenciones o estilos de palabras que las personas repiten al momento de relacionar un hipónimo con si hiperónimo dentro de un texto. Estos patrones si ya se encuentran registrados pueden permitirnos extraer instancias de la relación de hiponimia al aplicarse a un corpus.

Las primeras pruebas bajo patrones que se realizaron fueron construidas manualmente, es decir, después de observar la forma en la que los conceptos se describen y relacionan en un texto, un experto de dominio identificaba y formaba un conjunto de patrones sintácticos para crear una pareja hipónimo-hiperónimo.

En un trabajo realizado por Patrick Panel y Marco Pennacchiotti [12] se comenta que debido al reciente crecimiento de atención en problemas de enriquecimiento de conocimiento como responder preguntas de manera automática se ha motivado a los investigadores en procesamiento de lenguaje natural a desarrollar algoritmos para automáticamente buscar recursos semánticos. Con casi un sin fin de información textual a nuestra disposición, tenemos una grandiosa oportunidad para crecer de manera automática recursos ontológicos y bancos de datos. Su método es el siguiente:

Inducción de patrones En la fase de inducción de patrones, su algoritmo infiere un conjunto de patrones P que conecta a todas las instancias posibles dado un corpus. Cualquier patrón de aprendizaje funciona para esta etapa, se elije el mejor algoritmo y para cada instancia de entrada primero se obtienen todas las sentencias que contengan dos términos "X" y "Y", estas sentencias son generalizadas en un nuevo conjunto de sentencias reemplazando todas las expresiones terminológicas por una etiqueta terminológica. La generalización de términos es útil para pequeños conjuntos de documentos.

Clasificación y Selección de Patrones Un patrón confiable es aquel que es preciso y que puede extraer un numero mayor de instancias posibles.

Extracción de instancias En esta fase, se extraen las instancias "I" que coincidan con el patrón "P", a continuación se filtran las instancias incorrectas de acuerdo a un algoritmo propiedad de Patric & Marco [12].

En un conjunto de archivos o datos pequeños el numero de instancias extraídas puede ser demasiado pequeño como para garantizar suficiente evidencia o

entrenamiento para que en la siguiente iteración el algoritmo descubra de manera correcta instancias.

Cuentan con dos metodos para obtener instancias nuevas, vía web y vía sintáctica.

- Expansión Web: Nuevas instancias son extraidas de la web, usando el motor de búsqueda de Google, el sistema crea un conjunto de peticiones usando un patrón P instanciado con un concepto Y, por ejemplo, "Italia, País" y el patrón "Y como X", entonces la búsqueda en Google sera, "país como *", las instancias entonces son creadas de acuerdo al resultado de la búsqueda.
- Expansión Sintáctica: Nuevas instancias son creadas extrayendo expresiones correspondiendo a los términos mas importantes del texto.

Otro metodo por patrones propuesto por [5] aborda el problema de la extracción automática de parejas hipónimo-hiperónimo a partir de textos no estructurados tomados de la web. Su idea es formar un catálogo de hipónimos relacionado a un vocabulario predefinido y su método se basa en el uso de patrones. El método propuesto en su trabajo de investigación trata con patrones expresados en un nivel exclusivamente léxico su construcción es simple y no se necesita un fuerte conocimiento del idioma, no depende de analizadores sintácticos. Su trabajo considera como una pareja hipónimo-hiperónimo confiable y valida si es extraída en varias iteraciones o por varios patrones, y la confiabilidad de un patrón será mayor de acuerdo al número de parejas correctas que este recupere.

Descripción del método:

- Etapa 1: Descubrimiento de patrones mediante semillas.
- Etapa 2: Aplicar los patrones encontrados en la etapa uno y extraer tuplas de hipónimo-hiperónimo de una colección de documentos.
- Etapa 3: Ordenamiento de las tuplas obtenidas en la etapa dos, con el objetivo de ubicar las tuplas con mayor probabilidad de ser correctas en las primeras posiciones del catálogo.

3.4 Método de extracción de términos por medio de estadística.

Estos métodos estadísticos son aplicados para adquirir la relevancia de un termino para un dominio especifico. Un método estadístico popular es la frecuencia de un termino. El método de extracción de términos abordado en [13] usa la frecuencia de un termino en los documentos de corpus de dominio y el numero inverso de corpus en donde el termino aparece. Entre mas alta se la frecuencia del termino en comparación con la frecuencia de documentos en los que aparece mayor relevante es el termino.

Una vez que se obtienen los términos mas relevantes utilizando el método anterior, se procede a la formalizan de esos conceptos agrupándolos con sus atributos. Para derivar atributos de un corpus especifico se filtra y extraen las dependencias verbo/objeto y verbo/sujeto.

Para cada sustantivo que aparece en la frase que se esta analizando el verbo se utiliza como atributo para construir el contexto de la frase. Identificando

atributos similares de múltiples objetos conlleva a que las relaciones entre ellos puedan ser definidas. [14]

4 Conclusiones

La mayoría de los autores han realizado extracción automática de relaciones semánticas mediante patrones al ser uno de los métodos mas fáciles por tener ya reglas definidas para encontrarlas fácilmente, el problema es que puede no detectar relaciones que no respeten las reglas definidas previamente por los patrones que se acuerdan al principio por lo que puede ser un método no muy efectivo al no detectar todas las posibles relaciones.

Los métodos basados en diccionarios pueden ser una opción que parece viable y lo son, pero solo para clasificaciones generales, dado que los diccionarios abarcan un sin fin de temas, la taxonomía no sería de dominio lo que llevaría a un paso extra de crear o filtrar un diccionario a un tema en especifico para crear, detectar y limitar la taxonomía en un solo dominio.

El trabajo que se realizará a futuro sera basado en diccionario, se utilizara como corpus principal toda la base de datos de Wikipedia.org, el sistema podrá realizar dos tipos de consultas. 1.- Recibir un conjunto de palabras (animal, león) y el sistema deberá regresar que el león es un animal. 2.- Recibir una sola palabra (león) y el sistema deberá analizar en el corpus limitado que involucre a animales por ejemplo y deberá de identificar que leones un animal, y que puede tener también otras relaciones como que leones un mamífero. Atacaremos el problema de generalidad en diccionarios, limitando el diccionario a ciertos dominios en especifico.

References

1. Fellbaum, Christiane. WordNet and wordnets. In: Brown, Keith et al. (eds.), Encyclopedia of Language and Linguistics, Second Edition, Oxford: Elsevier, 2005
2. Mg. Augusto Cortez Vásquez, Mg. Hugo Vega Huerta, Lic. Jaime Pariona Quispe, Procesamiento de Lenguaje Natural. Facultad de Ingeniería de Sistemas e Informática Universidad Nacional Mayor de San Marcos. 2009
3. Julio Villena Román, Raquel M. Crespo García, José Jesús García Rueda. Procesamiendo del Lenguaje Natural. Universidad Carlos III de Madrid. 2012
4. Ríos Ríos, Aura Josefina; Bolívar, Constanza Ivet. Razonamiento verbal y pensamiento analógico. Universidad del Rosario. 2009
5. Rosa María Ortega Mendoza. Descubrimiento Automático de Hipónimos a partir de Texto no Estructurado. pages 23–27, 2007.
6. Alexander Panchenko, Sergey Adeykin, Alexey Romanov and Pavel Romanov. Extraction of Semantic Relations between Concepts with KNN Algorithms on Wikipedia. pages 78–88, 2012.
7. Ellen Riloff and Jessica Shepherd. A Corpus-Based Approach for Building Semantic Lexicons. pages 1–3, 1997 .
8. Philipp Cimiano, Andreas Hotho and Steffen Staab. Comparing Conceptual, Divisive and Agglomerative Clustering for Learning Taxonomies from Text. pages 1-2, 2004.
9. Rogelio Nazar, Jorge Vivaldi and Leo Wanner. Automatic taxonomy extraction for specialized domains using distributional semantics. 2012
10. Shachar Mirkin, Ido Dagan and Maayan Geffet. Integrating Pattern-based and Distributional Similarity Methods for Lexical Entailment Acquisition. page 2. 2006.
11. Carmen Klaussner and Desislava Zhekova. Lexico-Syntactic Patterns for Automatic Ontology Building. pages 1–3. 2011.
12. Patrick Pantel and Marco Pennacchiotti. Espresso: Leveraging Generic Patterns for Automatically Harvesting Semantic Relation. pages 1–3, 2006.
13. Kevin Meijer, Flavius Frasincar and Frederik Hogenboom. A Semantic Approach for Extracting Domain Taxonomies from Text. pages 1–5, March 2014.
14. Philipp Cimiano, Andreas Hotho and Steffen Staab. Learning Concept Hierarchies from Text Corpora using Formal Concept Analysis. page 2, August 2015.

Capítulo 4

Estado del arte en el poblado automático de ontologías.

Andrea Tamborrell, María Josefa Somodevilla

Benemérita Universidad Autónoma de Puebla, Blvd. 14 Sur y Av. San Claudio, Col. San Manuel, Ciudad Universitaria, Puebla, Pue., México.

andrea.tamborrell@anahuacxalapa.mx

Abstract. En el presente documento se realizó un estudio acerca de la situación actual en el procesamiento y poblado automático de ontologías, en base a trabajos e investigaciones realizados por distintos autores donde se abordaron metodologías desde la preparación de documentos de los cuales se realizará una ontología, obtención de conceptos que aparecen frecuentemente en dichos documentos, creación de propiedades, axiomas y restricciones hasta la creación de la ontología de manera automática en sí.

Keywords: ontologías automáticas, poblado automático de ontologías

1 Introducción

La extracción de información para instancias ontológicas es una tarea difícil, la cual involucra múltiples problemas a ser resueltos como:

- ¿Cómo identificamos la cadena de texto representando un sujeto de una sentencia?
- ¿Cómo encontramos el sentido de una cadena de texto y asignarla a la categoría semántica correcta?
- ¿Cómo extraemos valores de diferentes atributos del texto?

2 Metodología general de instanciación automática en una ontología

En [2] se dice que la extracción de información basada en ontologías es una disciplina en la cual el proceso de extracción de información de varios repositorios es guiado por una ontología. El proceso de extracción se lleva a cabo mediante múltiples pasos que pueden incluir el pre-procesamiento del texto en una máquina y definir heurísticas para identificar la información a ser extraída. La habilidad de extraer información de texto digital permite a diferentes aplicaciones como los sistemas de preguntas y respuestas dar respuestas más precisas.

Las ontologías se componen de información de dos tipos:

– Componentes como definición de clases, atributos y sus relaciones.
– Tripletas (Sujeto, objeto y predicado) que proveen información de las relaciones que existen en una instancia de la ontología y el sujeto y predicado asociado a cada una de esas relaciones.

Para poder identificar el dominio semántico del texto que se está procesando, los autores en [2] utilizan el concepto de lexicon semántico. Un lexicon semántico es un conjunto de palabras etiquetadas o relacionadas con una clase semántica, las cuales son parte integral del vocabulario de dominio, dichas palabras identifican plenamente un dominio. Básicamente un lexicon semántico es un diccionario con una red semántica de un mismo dominio.

Cada lexicon sintáctico está predefinido por expertos de cada dominio basado en su experiencia en ese dominio en particular.

El uso del lexicon semántico en el procesamiento de lenguaje médico se utilizó en [5] donde se ocupó el UMLS (Unified Medical Language System) el cual es un compendium de vocabulario de las ciencias biomédicas, donde cada palabra o lexema se asoció a una o más categorías sintácticas, de las cuales pueden tener más de una categoría sintáctica.

Un ejemplo más simple se encuentra en [2] donde se explica que un dominio de banco incluye palabras como: cuenta, cuenta de ahorros, pagos, transacciones y un lexicon sintáctico para un dominio de hotel incluye palabras como: desayuno, cena, habitación, servicio al cuarto.

Para la extracción de instancias de una ontologías en [2] se realizan los siguientes pasos:

1. Identificar temas o conceptos.
2. Identificar dominio de la ontología.
3. Extraer las clases de la ontología, atributos y valores.
4. Poblado de las instancias de la ontología.

Después de la instanciación de la ontología, como se menciona en [1], es necesario realizar actualizaciones periódicas. Para la instanciación automática de ontologías que de acuerdo a [1] se divide en cuatro fases secuenciales:

1. Fase de procesamiento del Lenguaje Natural y de procesamiento del corpus
2. Fase de reconocimiento e identificación de las entidades nombradas.
3. Fase de poblado de la ontología.
4. Verificación de la consistencia de la ontología.

Una vez que el dominio es identificado, el módulo extractor de instancias extrae la información de las instancias y actualiza la ontología.

3 Descripción de situación actual

El primer paso en el proceso de [2] es la identificación de la cadena más importante en el texto de la oración. Esta cadena forma la llave que básicamente hará

referencia al tema central de toda la oración. Por ejemplo si un párrafo de texto esta hablando de un hotel, entonces hotel es el tema central de todo el párrafo.

Una vez que se identifica el concepto principal del que se habla en el texto el siguiente paso es entender el dominio al que ese concepto pertence. Este paso es necesario para identificar el dominio apropiado de la ontología que se ocupará.

Por último se extraen los valores de varios atributos que son parte de la clase, una vez que estos valores son encontrados se agregan a la ontología. Dichos valores se extraen haciendo uso de la técnica de comparación de patrones, un patrón contiene unos cuantos términos y un conjunto de constantes en esos términos y cuando ese patrón es encontrado se toma el valor y es agregado a la ontología.

Abordando la metodología de otro autor, los pasos siguientes corresponden al método en [3] el cual es sencillo, rápido y una buena manera automática de obtener una organización inicial de conceptos de una colección de documentos. En este método es posible obtener una ontología que describe los conceptos de un documento individual o de una colección de documentos. Dicho método busca trabajar de manera automática haciendo que el usuario no se enfoque a la creación de ontologías pero es recomendable que revise el resultado final. Sus pasos son los siguientes:

- Preparación de documentos: En la primera fase, los documentos son preparados para obtener conceptos, si son textos largos se analiza la posibilidad de obtener un resumen, para documentos con estructuras ya definidas se trata de entender dicha estructura. Los resúmenes son leidos y se extraen los términos que serán utilizados en la preparación de la ontología.

- Obtención de conceptos: Inicialmente se obtienen los términos que aparecen en más del veinticinco porciento de los documentos, si este conjunto de términos es grande, puede ser reducido seleccionando un subconjunto de estos términos observando el criterio de quedarse con los terminos principales que aparezcan más frecuentemente en los documentos.

- Creación de propiedades, axiomas y restricciones: Al terminar la definición de conceptos en cada documento, se obtienen las relaciones semánticas para cada una de las ontologías, estas relaciones se organizan en dos propiedades, axiomas y restricciones. Hay muchas relaciones semánticas que pueden ser obtenidas usando Wordnet y es la herramienta más simple para la obtención de estas. Dejando a un lado Wordnet, los conceptos también son analizados por su grado de similitud, dependiendo de este valor los conceptos reciben una relación semántica de equivalencia.

- Creación de ontología: Esta es la última fase del método, el concepto y las relaciones semánticas son organizadas en ontologías que son almacenadas en archivos codificados en el lenguaje OWL. Este lenguaje es utilizado para definir ontologías y provee los mecanismos para la creación de componentes,

conceptos, instancias, propiedades y axiomas.

En [1] podremos encontrar diferentes metodologías para el poblado automático de ontologías, como lo son:

Metodología basada en la distancia contextual y la ganancia de conocimiento. Se basa en la distancia contextual, como elemento linguístico, y en la ganancia de conocimiento, como elemento ontológico. La distancia cotextual se refiere a la distancia física que hay entre dos unidades lingüísticas del texto. En cuanto a la ganancia de conocimiento es una medida que hace referencia a cuanto conocimiento se adquire en el sistema agregando ese elemento lingüístico a la ontología.

Metodología basada en roles semánticos. Los frames o roles semánticos son representaciones esquematizadas de situaciones del mundo real en base a las cuales se organiza la información. Las relaciones ontológicas y los frames son generalizaciones de situaciones cuya expresión lingüística consiste en una forma verbal. Todas aquellas relaciones ontológicas que contengan asociado un frame, contienen expresiones lingüísticas que representan dichas relaciones a un nivel textual y que describen cada relación.

Metodología de Gruninger y Fox Otra metodología escrita por Gruninger and Fox en [4] es en una primera instancia.

- Capturar escenarios de motivación: De acuerdo a ellos, el desarrollo de ontologías es motivado por escenarios que elevan el nivel de la aplicación. Los escenarios de motivación son problemas o ejemplos que no son atacados en las ontologías existentes.

- Formulación de preguntas informales de competencia: Son basadas en los escenarios obtenidos en el primer paso y pueden ser consideradas como requerimientos en forma de preguntas. Una ontología debe poder representar estas preguntas usando su terminología y ser posible de obtener más de una respuesta a esas preguntas haciendo uso de los axiomas.

- Especificación de la terminología de la ontología en un lenguaje formal

  - Obteniendo terminología informal: Una vez que las preguntas informales de competencia están disponibles, un conjunto de términos puede ser extraido de esas preguntas, estos términos servirán como base para especificar la terminología en un nivel formal de lenguaje.

    - Especificación formal de la terminología: Los conceptos se formalizan haciendo uso del "Knowledge Interchange Format" (KIF) que es un lenguaje diseñado para hacer uso de intercambio de conocimiento entre diferentes sistemas.

- Formulación de preguntas formales de competencia usando la terminología de la ontología: Una vez que se tiene el tercer paso completo y la terminología de la ontología ha sido definida, las preguntas de competencia son definidas formalmente.

- Especificación de axiomas y definiciones para los términos de la ontología en un lenguaje formal: Los axiomas de la ontología especifican la definición de términos y las restricciones en su interpretación. Los axiomas deben definir la semántica o significado de los términos a añadir a la ontología.

- Establecer condiciones para la creación de la ontología: Una vez que todo se formalizó se procede a la creación de la ontología con los términos extraídos, axiomas y preguntas.

Un ejemplo de aplicación que ha sido realizado con ontologías construidas en base a esta metodología es la siguiente:

Proyecto de agentes de integración de cadena de suministro: La meta es organizar la cadena de suministro como una red de cooperación, agentes inteligentes, cada uno realizando una o más tareas de cadena de suministro y cada uno coordinando sus acciones con otros agentes.

Otra metodología utilizada en el ámbito de la medicina se encuentra en [7]:

- Recolectar información que contenga el tema en específico de la ontología, desde libros, noticias y sitios web. Se realizó un estudio a fondo del tema y cuando se tenían dudas, dado que su trabajo sería aplicado a medicina, contactaban expertos como doctores, paramédicos y representantes médicos.

- Se iniciaba construyendo la ontología implementando conceptos que fueran parte de la ontología. Crearon una super clase llamada "Enfermedades transmitivas" y como subclases creaban las enfermedades en específico, sus síntomas y causas.

- Conceptualización: Buscar las relaciones entre los términos. Se realizó la implementación de relaciones binarias. (Influenza es una gripe), (gripe es transmitible) etc.

- Creación de instancias. Se crearon instancias individuales de conceptos que se diferenciaban a través de sus componentes.

Software para Poblado manual de una ontología Una ontología se puede realizar de manera manual con un software llamado Protégé, el cual es gratuito y de código libre para la construcción de modelos y aplicaciones basadas en conocimiento con ontologías.

En [6] se menciona que Protégé es la herramienta lider en la ingeniería de ontologías, tiene una arquitectura de software compleja, fácil de extender mediante plug-ins. Las ontologías son la base central de muchas aplicaciones como portales de conocimiento científico, sistemas de manejo e integración de información, comercio electrónico y servicios web.

Para la creación de ontologías en Protégé solo es necesario crear un nuevo proyecto y en una primera instancia empezar a crear las clases y subclases. Al tener clases y subclases es necesario definir las relaciones entre los conceptos o términos, esto se realiza mediante el lenguaje OWL y se agrega en un campo de texto en el software.

4 Conclusiones

Muchos de los autores que se mencionaron en este trabajo siguieron la metodología general de instanciación de una ontología siguiendo los pasos de primero identificar tema de la ontología, identificar los conceptos de interés, extraer las clases, atributos y valores y finalmente realizar el poblado de la ontología.
En los trabajos consultados se realizaron diferentes métodos para el pre-procesamiento del corpus, como fue procesamiento del lenguaje natural, prepararlos para obtener conceptos y en el caso de que sean muy extensos realizar un resumen, si el documento tiene una estructura se busca entenderla para que sea más fácil obtener la información más relevante.
Para poder identificar el tema de la ontología algunos autores se formulaban preguntas y posteriormente con los términos que se obtuvieron verificaban si con dichos términos podían responder a las preguntas formuladas.
Ya que se tiene definido el tema de la ontología es necesario consultar a fondo el tema y en el caso de que el tema de la ontología sea muy especializado en un área de la que no se tiene tanto conocimiento es recomendable consultar expertos en el área, además de personas que estén en contacto con dicha área.
Algunos autores utilizaron ciertas reglas para poder considerar un término como relevante para la ontología como es que aparezca en al menos un 25% de los documentos.
En varias de las metodologías mencionadas en el trabajo, al momento de definir los axiomas y restricciones se utilizaron herramientas para obtener las relaciones semánticas, se analizaron los conceptos para obtener un grado de similitud y de esta forma se puedan definir las relaciones semánticas de equivalencia.

References

1. Metodología para la población automática de ontologías. Aplicación en los dominios de medicina y turismo. Juana María Ruiz-Martínez. Procesamiento de Lenguaje Natural, Revista No.48, Marzo 2012.
2. Ontology Guided Information Extraction from Unstructed Text. Raghu Anantharangachar, Srinivasan Ramani, S Rajagopalan. International Journal of Web & Semantic Technology Vol. 4, No.1, January 2013.
3. Simple Method for Ontology Automatic Extraction from Documents. Andreia Dal Ponte Novelli, José María Parente de Oliveira. International Journal of Advanced Computer Science and Applications, Vol. 3, No.2, 2012.
4. Overview Of Methodologies For Building Ontologies. Fernández López, M.
5. A Semantic Lexicon for Medical Language Processing. Stephen B. Johnson. Journal of the American Medical Informatics Association, Vol. 6, No.3, May 1999
6. Model Driven Engineering and Ontology Development. Dragan Gasevic, Dragan Djuric, Vladan Devedzic, 2009.
7. Domain Ontology Development For Communicable Diseases. Iti Mathur, Hemant Darbari, Nisheeth Joshi, 2013.

Capítulo 5

Estado del Arte de Sistemas de Recuperación de Información*

Ana Laura Lezama Sánchez, Mireya Tovar Vidal y Darnes Vilariño Ayala

Benemérita Universidad Autónoma de Puebla,
Facultad de Ciencias de la Computación,
Avenida San Claudio, 14 sur,Ciudad Universitaria.
Puebla México.
{yumita1102,mireyatovar,dvilarinoayala}@gmail.com

Resumen En el presente trabajo se realiza una revisión de trabajos relacionados con Sistemas de Recuperación de Información (SRI) que incluyen la expansión semántica de consultas, con la finalidad de conocer los diferentes enfoques, métodos y herramientas utilizados por diversos autores en esta área de investigación. Cada autor citado, realiza la expansión semántica por medio de sinónimos, algoritmos genéticos entre otros, por lo general extraídos desde la ontología léxica WordNet, Ontologías de dominio etc. Por otro lado, se presenta también el estado del arte de SRI sin expansión de consultas, nuevamente para conocer sus enfoques, métodos y herramientas usadas por diversos autores. En base al análisis realizado se observa que los SRI con expansión obtienen mejores resultados dentro de sus experimentos, ya que cada sistema es capaz de traer mayor cantidad de infomación relevante al usuario que un SRI sin expansión debido a que solo es capaz de darle al usuario resultados exactos a su búsqueda.

Palabras clave Sistema de Recuperación de Información, expansión semántica de consultas, ontología léxica.

1. Introducción

La Recuperación de Información *(RI)* se centra en la representación, almacenamiento, organización y acceso a elementos de información [1]. Estos procesos deberían proporcionar al usuario la capacidad de acceder a la información que necesita. Sin embargo, existe un problema bastante importante en lo referente a la caracterización de las necesidades de información del usuario, que no suele ser fácil de solucionar [2].

La recuperación de información es el área de la ciencia y la tecnología que trata de adquirir, representar, almacenar, organizar y acceder a elementos de

* Esta investigación es parcialmente apoyada por el proyecto PRODEP-SEP ID 00570 (EXB-792) DSA/103.5/15/10854, por el proyecto ID 00570 VIEP-BUAP. Apoyado por el Fondo Sectorial de Investigación para la Educación, proyecto Conacyt 257357.

información [1]. Ya que, dada una necesidad de información del usuario, hecha en lenguaje natural, un sistema de RI produce como salida un conjunto de documentos cuyo contenido satisface potencialmente esa necesidad. Esta última observación es de suma importancia, ya que la función de un sistema de RI no es la de devolver la información deseada por el usuario, sino únicamente la de indicar qué documentos son potencialmente relevantes para dicha necesidad de información [1]. Dada la gran variedad de herramientas y métodos para la recuperación de información, consideramos que el estudio de los diferentes sistemas de recuperación con expansión y sin expansión, nos ayudará a conocer los avances y propuestas de cada autor, además de los resultados de cada investigación.

Este documento está estructurado de la siguiente manera, en la sección 2 se describe la información general de los SRI tradicionales en la literatura, así como los SRI con expansión de consultas y los SRI sin expansión, además de los trabajos relacionados por diversos autores en estos dos tópicos. Finalmente, en la sección 2.2 daremos nuestro punto de vista hacia las investigaciones estudiadas.

2. Sistema de Recuperación de Información

Los sistemas de recuperación de información, a menudo son comparados con las bases de datos relacionales. Tradicionalmente, los sistemas de recuperación de información, tienen información recuperada de textos no estructurados, lo que quiere decir, es que es texto en bruto, sin maquillaje. La diferencia fundamental entre bases de datos y sistemas de recuperación, es que las bases de datos son diseñadas para consultas de datos relacionales y que tienen conjuntos de archivos predefinidos, y los sistemas de recuperación de información, el modelo de recuperación, estructuras de datos y lenguajes de consulta [3].

A continuación se describen algunos de los tipos de sistemas de recuperación de información existentes.

Modelo booleano El Modelo de Recuperación Booleano *(MRB)* es uno de los métodos más utilizados para la recuperación de información. Este modelo se basa en la agrupación de documentos, los cuales están compuestos por conjuntos de términos y en la concepción de las preguntas como expresiones booleanas, de ahí deriva el nombre de modelo de recuperación booleano. Su característica principal es, que es un modelo de recuperación simple, basado en la teoría de conjuntos y el álgebra booleana. Se denomina Álgebra de Boole o Álgebra Booleana a las reglas algebraicas, basadas en la teoría de conjuntos, para manejar ecuaciones de lógica matemática. Se denomina así en honor de George Boole, famoso matemático, que la introdujo en 1847. Dada su inherente simplicidad y su pulcro formalismo, ha recibido gran atención y ha sido adoptado por muchos de los primeros sistemas bibliográficos comerciales. Su estrategia de recuperación está basada en un criterio de decisión binario (pertinente o no pertinente) sin ninguna noción de escala de medida, sin noción de un emparejamiento parcial en las condiciones de la pregunta [3].

Modelo Espacio Vectorial El Modelo de Recuperación Vectorial o de Espacio Vectorial *(MEV)* propone un marco en el que es posible el emparejamiento parcial a diferencia del modelo de recuperación booleano, asignando pesos no binarios a los términos índice de las preguntas y de los documentos. Estos pesos de los términos se usan para calcular el grado de similitud entre cada documento guardado en el sistema y la pregunta del usuario. Su característica principal es que, ordenando los documentos recuperados en orden decreciente a este grado de similitud, el modelo de recuperación vectorial toma en consideración los documentos que sólo se emparejan parcialmente con la pregunta, así el conjunto de la respuesta con los documentos alineados es mucho más preciso (en el sentido que empareja mejor la necesidad de información del usuario) que el conjunto recuperado por el modelo booleano. Los rendimientos de alineación del conjunto de la respuesta son difíciles de mejorar. La mayoría de los motores de búsqueda lo implementan como estructura de datos y el alineamiento suele realizarse en función del parecido (o similitud) de la pregunta con los documentos almacenados [3].

Modelo probabilístico El Modelo de Recuperación Probabilístico *(MRP)*, la base principal de su funcionamiento es el cálculo de la probabilidad de un documento de ser relevante a una pregunta dada. Los modelos anteriores están basados en la equiparación en la forma más "dura". En el booleano es o no coincidente, y en el vectorial el umbral de similitud es un conjunto, y si un documento no está no es similar y, por lo tanto, no recuperable. Su característica principal es la equiparación probabilística se basa en que, dados un documento y una pregunta, es posible calcular la probabilidad de que ese documento sea relevante para esa pregunta [3].

2.1. Sistemas de Recuperación de Información con Expansión de Consultas

La expansión de cosultas o *(Query Expansion)* es la técnica comunmente usada en Recuperación de Información, para mejorar el desempeño de los resultados por reformulación de la consulta original, ya sea añadiendo nuevos términos o reponderación de los términos originales [4]. Los términos de la expansión de consultas pueden ser automáticamente extraídos de los documentos, o tomándolos de recursos de conocimiento, como tesauros, ontologías léxicas como WordNet [5,6], algoritmos genéticos, etc. La ventaja de dichas técnicas es la expansión de términos que son extraídos de la colección [4].

A continuación se describen algunos trabajos relacionados con Sistemas de Recuperación de Información que utilizan expansión de consultas.

En Cotelo et. al. [7] presentan la propuesta de un SRI donde dada una consulta, retorne un conjunto de documentos relacionados. En su sistema implementado, definieron un lenguaje de consulta semántico llamado SemQL, que tiene como objetivo acortar las distancias entre la consulta ingresada por el usuario y los documentos indexados, después realizan un análisis de dependencias,

y la implementación de un motor de búsqueda para recuperar documentos, la búsqueda es mediante palabras claves originarias de la consulta, obtienen el conjunto de documentos relevantes, en primera instancia, enfocándose más en una búsqueda semántica. Para la recuperación de documentos los autores usaron Lucene+SOLR para la que generaron una interfaz que permitió indexar y recuperar documentos comunicándose por JSON, hicieron uso de la base de datos léxica WordNet y aplican expansión de consultas por sinónimos. Usarón la herramienta Django para la implementación de la interfaz de usuario. Generan un corpus llamado BusSem-2012, cuyo contenido son artículos obtenidos del sitio web del diario británico *The Times*. De acuerdo a sus experimentos realizados llegaron a la conclusión de que si el algoritmo propuesto fuera implementado en los buscadores como Google, Yahoo, Altavista, etc., la experiencia de la búsqueda del usuario mejoraría notoriamente.

En Kuna et. al. [8] en su investigación cobran mayor relevancia los metabuscadores (uno de los principales modelos de SRI que operan sobre internet). Presentan un método de expansión de consultas para un SRI (metabuscador) basado en la utilización de una ontología de dominio específico. Para lograrlo, diseñaron su propia ontología dentro de la sub-área de Inteligencia Artificial, utilizando la herramienta Protégé [9] que es una herramienta de software específica para operar con ontologías. El algoritmo desarrollado para la expansión de consultas, se compone de dos etapas, por un lado una primera instancia de selección del concepto de la ontología más similar a la consulta original y luego la generación de las consultas. Como resultado de la ejecución de las dos fases del algoritmo se cuenta con una serie de cadenas de caracteres correspondientes a las expansiones de la consulta original.

En Schneider et. al. [10] plantean el uso de una ontología para mejorar los resultados de búsqueda en un SRI en un dominio en particular. Fue desarrollada dentro del marco de su investigación, que se centró en el dominio financiero. En el que se distinguen dos capas, la primera que relaciona todas la entidades presentes en el mercado bursátil y la segunda que asigna metadatos a cada una de las entidades. La expansión de la consulta fue abordada desde dos puntos de vista, lanzando la consulta completa del usuario en lenguaje natural y el análisis semántico de la consulta para expandir únicamente las entidades. El análisis semántico de la consulta se realiza utilizando Textalytics [1]. La búsqueda en lenguaje natural se basa en la utilización conjunta de la ontología y Lucene. Para la evaluación de la búsqueda en la ontología, diseñaron la prueba de Cranfield y la evaluación de cada una de las consultas es por medio de precisión y recuerdo.

En Shabanzadeh et. al. [11] presentan un algoritmo de expansión de consultas basado en relaciones semánticas. Usan WordNet para extraer relaciones semánticas entre palabras como sinónimos, hiperónimos e hipónimos. El algoritmo propuesto primero retira ambigüedad y palabras vacías presentes en la consulta y extrae las palabras claves existentes, entonces agrupan las palabras de la consulta basándose en su similitud semántica, y extrayendo palabras relacionadas a todo el grupo de palabras en lugar de cada palabra de la consulta.

[1] http://textalytics.com

Después construyen una red semántica de palabras de la consulta y las palabras relacionadas a la consulta, en esta red, las palabras son los nodos y cada nodo es vinculado a nodos que están relacionados semánticamente. Analizaron los documentos recuperados calculando precisión y recuerdo, usaron el modelo espacio vectorial para recuperar documentos, además de un conjunto de datos del archivo *SMART* del departamento de Ciencias de la Computación de la Universidad de Cornell llamado *TIME* que es una colección de 1963 de noticias que contiene 425 artículos pequeños y 80 consultas que se encuentran en lenguaje natural.

En Soni et. al. [12] proponen un algoritmo genético para la expansión de consultas hechas en lenguaje natural, se utiliza el coeficiente de Czekanowski durante el proceso de expansión para que la recuperación sea más eficiente, ya que mide la similitud entre los documentos recuperados y la consulta dada. Utiliza un analizador de texto que ayuda a encontrar palabras claves en los documentos, que serán utilizadas para hacer el cromosoma que es la base del algoritmo genético. La expansión de la consulta, es realizada en base al documento que tenga el cromosoma con el mejor valor en su función de aptitud, para después hacer la expansión manualmente, usando una medida de similitud. Observaron que el uso de un algoritmo genético aumenta la relevancia de documentos recuperados, si la tasa de mutación es menor el cromosoma converge en una sola generación.

En Harb et. al. [13] usaron un rastreador que debe recorrer la WWW para obtener documentos en el dominio del cuidado de la salud, específicamente en enfermedades ictéricas. El modelo espacio vectorial es adaptado en la propuesta de este trabajo para la representación de documentos, retira palabras vacías, etc. La consulta es expandida por sinónimos extraídos de Wordnet, pero solo con aquellos sentidos más comunes de cada término de la consulta. Con el método de recuperación de información semántico propuesto se han aprovechado las ventajas de la web semántica para recuperar documentos pertenecientes al dominio mencionado. Supera el método de recuperación de información clásica y demuestra mejoras en el rendimiento.

En Mahgoub et. al. [14] introducen una aproximación de expansión de consultas usando una ontología construída con páginas de wikipedia, además de otros tesauros para mejorar la precisión en la búsqueda del idioma árabe. Su aproximación, depende de tres recursos árabes que son Wikipedia en árabe, como el recurso con mayor información semántica, el diccionario Al Raed, que es un diccionario monolingüe para palabras modernas, y el diccionario Google_WordNet que es una colección de todas las palabras en WordNet y traducidas con el traductor de Google. La indexación y recuperación de su sistema depende de Lucene. Para expandir la consulta, primero localizan el nombre de las entidades o conceptos que aparecen en la consulta , si el nombre de una identidad o concepto es localizado, agregan el título de redirigir la página que conduce al concepto similar agregando una subcategoría del sistema. Para sus experimentos, usaron el conjunto de datos construídos desde el libro "Zad Al Ma'ad", dicho conjunto de datos contiene 25 consultas y 2,730 documentos.

En Fernández et. al. [15] muestran aspectos relacionados con la integración de la tecnología disponible del tratamiento del lenguaje natural en el desarrollo

de un metabuscador que alcance un mayor grado de acierto en la recuperación de información realizada por un buscador tradicional así como en el tratamiento posterior de los documentos recuperados. Describen su proceso realizado para la expansión de las consultas de los usuarios con información lingüistica empleando dos recursos léxicos para el castellano: ARIES que es un léxico morfológico desarrollado por la Universidad Politécnica de Madrid y la Universidad Autónoma de Madrid para el tratamiento de la morfología y EuroWordnet [16] para el tratamiento de la semántica. La generación de la consulta está compuesta por dos tareas principales, la primera consiste en transformar la consulta del usuario en lenguaje natural (LN) en una consulta formal que el buscador pueda ejecutar. La segunda funcionalidad consiste en extender los términos significativos de la consulta (formal) utilizando conocimiento lingüistico; para ello se añaden a los términos significativos de la consulta (enlazados con AND) las variantes morfológicas y semánticas mediante OR con el fin de construir una consulta en forma normal conjuntiva. Su trabajo forma parte del sistema MESIA, Modelo computacional para extracción selectiva de información de textos cortos, que amplía la búsqueda habitual (consulta y presentación de resultados) con nuevas capacidades morfológicas y semánticas y analiza otros aspectos obtenidos a partir de la estructura de las páginas, del tratamiento lingüístico de algunas de las unidades de texto seleccionadas automáticamente y de la experiencia de uso.

En Valbuena et. al. [17] la metodología que siguen es definir en primer lugar la estructura del sistema de búsqueda, que está compuesta por el módulo de procesamiento de consultas y el módulo de emparejamiento-ranking. Su sistema desarrollado en java esta soportado en un sistema de indexación ontológico, por lo tanto, eligieron las ontologías Cell Type Ontology(CL) y Protein Ontology (PO), ambas pertenecientes al corpus *CRAFT*. Los sinónimos son extraídos directamente de la ontología, dichos sinónimos son utilizados durante el preprocesamiento de consultas, las cuales son expandidas usando ontologías, utilizan la técnica de emparejamiento entre cadenas de Levenstein (haciendo uso de programación dinámica). Los autores se basan en el modelo espacio vectorial e hicieron uso de la biblioteca Hadoop que es una implementación abierta de MapReduce. Utilizan Jena, que provee una serie de herramientas para construir aplicaciones de la web semántica y posibilita la realización de consultas a ontologías. Además utilizan *Lucene Analyzer* (de *Apache Software Foundation*) para encontrar las raíces de los términos, ya que contiene la clase EnglishStemmer que implementa la versión inglesa del algoritmo de porter. El proyecto fue evaluado con la métrica de precisión.

En Cruanes et. al. [18] proponen una aproximación de mapeado de información en lenguaje natural del dominio de enfermería, utilizando métodos de similitud léxica. Los autores generan expansión por sinónimos y buscan antonímia. No usaron recursos como EuroWordNet, ya que de acuerdo a los autores, no se ajustaba a las necesidades del dominio estudiado.

En Deco et. al. [19] proponen un refinamiento semántico, que guiará al usuario a desambiguar los términos ingresados por el. Realizaron expansión semántica

de consultas por sinónimos, usaron WordNet, y en la generación de la estrategia, ocuparon operadores lógicos.

En la tabla 1 se presenta un resumén de los trabajos revisados anteriormente. En la Tabla se obseva los recursos léxicos, los dominios, el tipo de expansión y el sistema de recuperación de información que cada autor usó en su investigación.

Tabla 1. Estado del arte de Sistemas de Recuperación de Información con expansión de consultas

Autores	Recurso lexicos	Dominios	Expansión de consulta	Tipo SRI
[7]	WordNet	Noticias	sinónimos	Lucene
[8]	Protègè	IA	Ontología de dominio	Metabuscador
[10]	Textalytics	Financiero	Análisis semáticos	Lucene
[11]	WordNet	Noticias	Relaciones semáticas	MEV
[12]	Analizador de texto	-	algoritmo genético	MEV
[13]	Analizador de texto	Cuidado de salud	sinónimos	MEV
[14]	WordNet	Idioma Árabe	Ontología de dominio	Lucene
[15]	ARIES/EuroWordNet	Festivales	sinonimia/hiponimia	Lucene
[17]	SPARQL	Cuidado de la salud	sinonimia/Ontología	Lucene
[18]	Métodos de similitud léxica	Enfermería	sinónimos	-
[19]	WordNet	Cuidado de la salud	sinónimos	Booleano

2.2. Sin Expansión de Consultas

Los SRI sin expansión de consultas consisten en que el usuario plasma su necesidad de información en una consulta aceptada por un SRI, por su parte el SRI transformará dicha consulta en una representación interna que permita su comparación con los documentos indexados. La consulta supone un intento por parte del usuario de especificar las condiciones que permitan acotar dentro de la colección aquel subconjunto de documentos que contienen la información que desea. Por lo tanto, el SRI parte de la consulta formulada por el usuario, no de la necesidad de información original, por lo que una formulación incorrecta o insuficiente no podrá guiar adecuadamente al SRI durante el proceso de búsqueda. A este respecto los mayores problemas a los que ha de hacer frente el SRI son, por una parte, la escasa habilidad del usuario a la hora de formular su necesidad en forma de consulta y, por otra parte, que a la hora de describir un mismo concepto los términos empleados por el usuario y los autores de los documentos suelen diferir, impidiendo el establecimiento de correspondencias [1].

Sistemas de Recuperación de Información sin Expansión de Consultas En Tovar [20] se utiliza un sistema de recuperación de información, para la evaluación de la ontología, usando los términos de los conceptos y así recuperar documentos relevantes a los mismos sin utilizar expansión. Como observación final, el autor, considera que el uso de la expansión podría mejorar los resultados de la evaluación global de la ontología de dominio.

En Gil et. al. [21] se propone el desarrollo de un SRI en Inteligencia Artificial enfocada a textos médicos con el objetivo de conseguir un sistema destinado a introducirse en el campo de la medicina personalizada y el campo turístico. Usa la base de conocimiento colaborativa Freebase [22] para recuperar listas de conceptos médicos o turísticos y conectarlos con contenidos semánticamente relacionados. Además de recursos formales como PubMed que es usado para obtener casos científicos, ya que es una de las fuentes más utilizadas para buscar literatura biomédica (1300 millones de búsquedas en 2009) y MedlinePlus.

En Cisneros et. al. [23] se presenta un prototipo que busca informaciíon sobre términos médicos en colecciones divulgativas de medicina multilingües (español, inglés, árabe y japonés). La arquitectura del sistema consta de un módulo de búsqueda, la cual para su creación usa la herramienta IR Apache Lucene [24] que es un módulo de extracción de información que permite procesar textos ya sean resultados del módulo de búsqueda o introducidos por el usuario. Los textos son analizados semánticamente por MetaMap que realiza el análisis sintáctico de cualquier texto biomédico así como la detección de conceptos del metatesauro Unified Medical Language System en él. Emplean el sistema DrugDDI para la detección de interacciones entre fármacos en textos biomédicos que están basados en determinar automáticamente los patrones que identifican interacciones entre fármacos de un conjunto de documentos [25] y finalmente el sistema busca información para cada fármaco detectado.

En Hernández et. al [26] se desarrolló un prototipo que consta de una interfaz web que permite la búsqueda y visualización de resultados a partir de una consulta dada. En dicho sistema de búsqueda textual se realiza un pre proceso con la información textual extraída de los documentos multimedia, se realiza un análisis de detección de las entidades nombradas, una indexación que consiste en la creación de un único índice haciendo uso de Lucene [24], y una vez indexado el corpus Deportes [27] que es una colección compuesta por 4 tipos de recursos o documentos multimedia; en la búsqueda de cada consulta se obtendrá una única lista de resultados ordenados por relevancia. Lucene se basa en los modelos vectorial y booleano puro e incluye la posibilidad de incluir entre sus consultas operadores booleanos, búsquedas basadas en campos, etc.

En Artiga et. al. [28] proponen un modelo de evaluación de Interfaces en SRI, mediante el análisis de diversos aspectos encontrados en cada SRI evaluado. Desarrollaron un análisis de las diversas operaciones y procesos que se llevan a cabo en los SRI, obtuvieron un modelo de comparación y evaluación que les permitió identificar diferentes aspectos.

En Herrera et. al. [29] presentan un SRI definido usando un enfoque lingüistico difuso ordinal, además de estar basado en un lenguaje de consultas Booleano ponderado . El SRI es un operador or-and del operador LOWA, que usaron para agregar información lingüistica ordinal. El operador LOWA es aplicado para modelar la evaluación de los conectivos lógicos AND y OR durante la evaluación de consultas, es así como el concepto de soft o computo flexible se introduce en la recuperación del SRI.

En la tabla 2 se presentan los recursos léxicos que cada autor utiliza en su investigación, así como los dominios con los que trabajan y el tipo de SRI que utilizan.

Tabla 2. Estado del arte de Sistemas de Recuperación de Información sin expansión de consultas

Autores	Recurso lexicos	Dominios	Tipo SRI
[20]	-	IA/EOR/SCORM/TURISMO	Booleano
[21]	PUBMED/MEDLINEPLUS	IA/TURISTICO	GATE
[26]	Stilus/Freeling/Snowball	Deportes	Lucene
[29]	LOWA	-	Booleano ponderado

Conclusiones

Se han revisado los diferentes trabajos con expansión de consultas, de diversos autores, donde usan recursos léxicos como WordNet, Protègè, analizadores léxicos, métodos de similitud léxica entre otros. Además de dominios como noticias, IA, financiero, cuidado de la salud, etc. en su mayoría cada autor realiza expansión por sinónimos, pero también realizan expansión por relaciones semánticas, hipónimos, etc. y los tipos de SRI empleandos son el modelo espacio vectorial y Lucene, etc. Para los trabajos que no realizan expansión de consultas los recursos léxicos empleados, fueron Freeling, Snowball, LOWA, PUBMED, etc, los dominios usados, deportes, IA, EOR, SCORM y Turismo, y los tipos de SRI empleados fueron Booleano, GATE y Lucene. Después de que se han revisado los diferentes trabajos de SRI con expansión y sin expansión, hemos llegado a la conclusión de que la la expansión de los términos de la consulta mejora notablemente el rendimiento de los sistemas. Puesto que se incrementa la cantidad de información que puede ser relevante al usuario en su proceso de búsqueda. Sin embargo, los sistemas sin expansión solo se limitan a la búsqueda de las palabras que integran a la consulta misma limitando la información que puede ser útil al usuario.

Referencias

1. Jesús Vilares Ferro. *Aplicación del procesamiento del lenguaje natural en la recuperación de información en español*. PhD thesis, Universidad da Coruña, Departamento de Computación, Mayo 2005.
2. Manuel Blázquez Ochando. Técnicas avanzadas de recuperación de información: Procesos, técnicas y métodos. 2013.
3. Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.
4. Olga Vechtomova and Ying Wang. A study of the effect of term proximity on query expansion. *Journal of Information Science*, 32(4):324–333, 2006.

5. George A Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.
6. Christiane Fellbaum. A semantic network of english verbs. *WordNet: An electronic lexical database*, 3:153–178, 1998.
7. Santiago Cotelo, Alejandro Makowski, Luis Chiruzzo, and Dina Wonsever. Búsqueda de documentos utilizando criterios semánticos. 2012.
8. Horacio Daniel Kuna, Martín Rey, Lucas Podkowa, Esteban Martini, and Lisandro Solonezen. Expansión de consultas basada en ontologías para un sistema de recuperación de información. In *XVI Workshop de Investigadores en Ciencias de la Computación*, 2014.
9. Holger Knublauch, Ray W. Fergerson, Natalya F. Noy, and Mark A. Musen. The protégé owl plugin: An open development environment for semantic web applications. pages 229–243. Springer, 2004.
10. Julián Moreno Schneider, Thierry Declerck, José Luís Martínez Fernández, and Paloma Martínez. Prueba de concepto de expansión de consultas basada en ontologías de dominio financiero. *Procesamiento del lenguaje natural*, 51:109–116, 2013.
11. Mozhgan Shabanzadeh, Mohammad Ali Nematbakhsh, and Naser Nematbakhsh. A semantic based query expansion to search. In *Intelligent Control and Information Processing (ICICIP), 2010 International Conference on*, pages 523–528. IEEE, 2010.
12. Neha Soni and Jaswinder Singh. Relevancy enhancement of query with czekanowski coefficient by expanding it using genetic algorithm. 2011.
13. Hany M Harb, Khaled M Fouad, and Nagdy M Nagdy. Semantic retrieval approach for web documents. *International Journal of Advanced Computer Science and Applications (IJACSA)*, 2(9):11–75, 2011.
14. Ashraf Y Mahgoub, Mohsen A Rashwan, Hazem Raafat, Mohamed A Zahran, and Magda B Fayek. Semantic query expansion for arabic information retrieval. In *EMNLP: The Arabic Natural Language Processing Workshop, Conference on Empirical Methods in Natural Language Processing, Doha, Qatar*, pages 87–92, 2014.
15. Paloma Martínez Fernández and Ana García Serrano. Utilizando recursos lingüísticos para mejora de la recuperación de información en la web. *Inteligencia Artificial, Revista Iberoamericana de Inteligencia Artificial*, (16):55–64, 2002.
16. Piek Vossen. Introduction to eurowordnet. In *EuroWordNet: A multilingual database with lexical semantic networks*, pages 1–17. Springer, 1998.
17. Sonia Jaramillo Valbuena and Jorge Mario Londoño. Búsqueda de documentos basada en el uso de índices ontológicos creados con mapreduce. *Ciencia e Ingeniería Neogranadina*, 24(2):57, 2014.
18. Jorge Cruanes, M Teresa Romá Ferri, and Elena Lloret Pastor. Análisis del uso de métodos de similitud léxica con conocimiento semántico superficial para mapear la información de enfermería en espaol. *Procesamiento del lenguaje natural*, 49:75–82, 2012.
19. Regina Motz, Claudia Deco, Cristina Bender, Jorge Saer, and Mario Chiari. Refinamiento semántico para recuperación de información desde la web. In *Proceedings Workshops on Artificial Intelligence, Iberamia*, pages 172–179, 2004.
20. Tovar Vidal Mireya. *Evaluación automática de ontologías de dominio restringido*. PhD thesis, Cenidet, Febrero 2015.
21. Rafael Muñoz Gil, Fernando Aparicio, and Manuel de Buenaga. Sistema de acceso a la información basado en conceptos utilizando freebase en español-inglés sobre el dominio médico y turístico. *Procesamiento del lenguaje natural*, 49:29–38, 2012.

22. Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM, 2008.
23. Daniel Sánchez Cisneros, Isabel Segura Bedmar, and Paloma Martínez Fernández. Prototipo buscador de información médica en corpus multilingües y extractor de información sobre fármacos. *Procesamiento del Lenguaje Natural*, 49:209–212, 2012.
24. Branko Milosavljevic, Danijela Boberic, and Dušan Surla. Retrieval of bibliographic records using apache lucene. *The Electronic Library*, 28(4):525–539, 2010.
25. Sandra García-Blasco, Roxana Danger, and Paolo Rosso. Drug-drug interaction detection: a new approach based on maximal frequent sequences. *Procesamiento del lenguaje natural*, 45:263–266, 2010.
26. David Hernández-Aranda, Rubén Granados, and Ana García-Serrano. Servicios de anotación y búsqueda para corpus multimedia. *Procesamiento del Lenguaje Natural*, 49:213–216, 2012.
27. Chauhan R., Goudar R., Rathore R., Singh P., and Rao S. Expansión de consultas automática basada en ontología para la recuperación de información semántica en dominio deportivo. *ICECCS*, 2012.
28. Viviana Natividad Asensi Artiga and Juan Antonio Pastor Sánchez. Un modelo para la evaluación de interfaces en sistemas de recuperación de información. In *La representación y la organización del conocimiento en sus distintas perspectivas: su influencia en la recuperación de la información: actas del IV Congreso ISKO-España EOCONSID'99, 22-24 de abril de 1999, Granada*, pages 401–409. Universidad de Granada, 1999.
29. E Herrera-Viedma and J Domínguez. Un sri lingüístico difuso basado en el operador lowa. 2012.

Capítulo 6

Modelo computacional en apoyo a trastorno específico del lenguaje

José Abraham Baez Bagatella

Benemérita Universidad Autónoma de Puebla

Abstract. El trastorno específico del lenguaje (TEL) es caracterizado por problemas en hablar, comprender o expresarse verbalmente por medio del idioma, notandose principalmente mediante el idioma hablado, algunos autores lo nombran como Trastorno Específico del Desarrollo del Lenguaje (TEDL). Este trastorno impide a una persona el facil y correcto aprendizaje de la lengua madre, comenzando a tener problemas a edades cortas para empezar a expresarse verbalmente, las personas con este trastorno sufren de deterioros constantes y a gran escala de la comprension y expresion del vocabulario, gramatica, contexto y semantica mediante el uso hablado y escrito del idioma.

Keywords: trastorno, específico, lenguaje, dislexia, autismo, déficit, hiperactividad, atención, fonología, fonogramas, gramática, semántica, retraso, psicologia, neurologia, modelo, computacional, apoyo, niños

1 Generalidades del trastorno

La definición que se ha adoptado parte de unos criterios de exclusión previamente fijados y que se han aceptado universalmente: "todo inicio retrasado y todo desarrollo lento del lenguaje que no pueda ponerse en relación con un déficit sensorial (auditivo) o motor, ni con deficiencia mental, ni con trastornos psicopatológicos (trastornos masivos del desarrollo en particular), ni con privación socioafectiva, ni con lesiones o disfunciones cerebrales evidentes"[7].

El trastorno específico del lenguaje (TEL) es caracterizado por problemas en hablar, comprender o expresarse verbalmente, tales problemas comunmente son fallas en la escucha y reinterpretación de la información recibida, también hay deficiencia en la expresión oral del idioma y ademas fallas en la correcta lectura y comprensión de lo que se lee. Este trastorno se presenta desde edades cortas y es altamente dificil de detectar pues podria confundirse o solaparse con otros trastornos que se relacionan con el trastorno específico del lenguaje. Este sindrome o trastorno esta relacionado con diversas condiciones como: daño cerebral, autismo, dislexia, deficit de atención, hieractividad entre otras, mas también puede ocurrir en personas que no presentan alguna de las condiciones mencionadas, es decir personas con una aparente condición "normal".

Un particular interés ha ocurrido en la relación que tienen el síndrome de déficit de atención y el trastorno específico del lenguaje debido a la carencia de lectura, se ha establecido que los niños con déficit de atención presentan carencia de lectura de 25% a 40%, y los niños con trastorno específico del lenguaje presentan carencia de lectura de un 25% a 30%. Si se prueba la existencia una forma pura de TEL entonces se podría asumir que existe un subsistema en el cerebro que controla la gramática. Algunos análisis a nivel genético indica que estos dos trastornos se enlazan sobre la misma localización genética (6p22). Este estudio preliminar podría indicar que ambos síndromes se transmiten de manera genética.[4]

No ha sido posible determinar a que edad se completa el desarrollo de la percepción del habla, pero se ha encontrado que en bebes de 1 a 4 meses logran distinguir el sonido como lo hace una persona adulta, así como que la categorización de los sonidos muestra mejores índices a los 4 que a los 7 años y finalizando con mejoras en los contrastes auditivos entre los 9 y 17 años, lo cual sugiere una tendencia evolutiva.[6]

No debe confundirse un trastorno del lenguaje con otras condiciones que parecen serlo, por ejemplo la falta de inteligencia causado por daño cerebral puede parecer que suma un trastorno específico de lenguaje, pero no es siempre la situación, la detección de este trastorno específico se vuelve aun más dificil en personas con daño cerebral.

Como se ha mencionado anteriormente, no es necesario sufrir de alguna de estas condiciones para padecer TEL. Existe un debate para determinar cual es la causa de este trastorno, por una parte existe la teoría que establece que es un déficit específico para el lenguaje y por otra parte la teoría que afirma que es causado por limitaciones de procesamiento en general.

La primera teoría asume que los niños con TEL tienen dificultades para aprender el habla desde una edad muy temprana y eventualmente se estancan en el aprendizaje, este estancamiento se puede reflejar en otros sistemas cognitivos conforme trascurre el tiempo[2].

La segunda teoría afirma que este síndrome no es culpa del material lingüístico, sino como es procesado dentro del cerebro, además se piensa que algunos trastornos de carácter no lingüístico tales como la memoria o la percepción también pueden ser responsables de este desorden. La teoría mas prominente de este tipo es la "Hipótesis del Déficit Rápido de Procesamiento Temporal" la cual sostiene que el TEL es resultado en el déficit en procesar de manera breve o rápida la información que se escucha [2].

Además se han encontrado mediante estudios problemas en las funciones motoras de niños con autismo que guardan estrecha relación con el TEL.

2 Trastorno específico del lenguaje en niños

Para la mayoría de los bebes y niños la comprensión del lenguaje se desarrolla naturalmente desde su nacimiento, se activan sus capacidades de recuerdo e imitación, por medio de la vista o el oído, con lo cual comienzan a entender el (los) idioma(s). Sin embargo se estima que 1 en cada 20 niños tiene un síndrome de trastorno del lenguaje, el cual si se desconoce su causa se le denomina trastorno de desarrollo del lenguaje[1].

Los niños que presentan este trastorno tienen como síntomas: déficit en varios aspectos del lenguaje tales como la fonología, morfología y la sintaxis, además mala identificación de los tiempos de conjugación y de los artículos. Por ejemplo niños con lengua materna Inglés pueden hacer malas sustituciones de sonidos, tales como omitir el sonido de las consonantes al final de una palabra o bien reemplazar sonidos parecidos en la forma de pronunciar una consonante.

Es difícil determinar este síndrome pues hasta la edad de 4 años un niño puede tener las mismas capacidades de lenguaje como niños de entre 5 y 6 años, sin embargo en algunos casos puede ocurrir un estancamiento del lenguaje que perdura hasta la edad adulta.[3]

3 Condiciones que presentan TEL.

3.1 Dislexia

La dislexia es una enfermedad especifica de aprendizaje que se presenta en niños que han sido instruidos de manera adecuada y poseen una inteligencia "normal", las dificultades de aprendizaje que presentan son debido a una deficiencia en su forma de procesar lo que se escucha, lo que obstaculiza la comprensión y el correcto uso de la gramática del idioma. El déficit fonológico en las personas con dislexia ha sido demostrado por distintos equipos de investigación en diferentes lenguas y utilizando una amplia variedad de tareas.[5]

Los niños con dislexia presentan un retraso constante en el desarrollo de la percepción del habla, esto durante la educación básica. Las habilidades de discriminación de las claves fonéticas necesarias para percibir el habla en los niños con dislexia quienes sólo avanzan en la discriminación del punto de articulación[5]. Además hay una carencia para procesar el ritmo del habla y la temporalidad de la misma. Sin embargo presentan una buena habilidad para distinguir entre las variantes acústicas del mismo fonema, por ejemplo "dado", les es mas fácil diferenciar la acústica de la primera "d" con respecto a la segunda.

Los estudios enfocados en niños con dislexia que examinan las dificultades en el proceso perceptivo auditivo plantean dos hipótesis: "la hipótesis del déficit de procesamiento temporal y la hipótesis del déficit específico en percepción del habla"[5]. El déficit de procesamiento temporal altera la manera en que se

perciben los sonidos de corta duración y se reflejan en problemas para procesar sonidos breves, por ejemplo los pronombres "El", "La", etc. La segunda hipótesis plantea que las dificultades de lectura surgen a partir de un problema de percepción y procesamiento auditivo para diferenciar las representaciones fonológicas de los fonemas.

Teniendo en cuenta lo anteriormente abordado podría llegar a decirse que la dislexia es resultado de un retraso en el desarrollo del lenguaje, pero las características acústicas relevantes varían conforme al desarrollo de la persona, la experiencia y la edad. Sin embargo los datos sobre esta condición aun son escasos.

3.2 Trastorno por déficit de atención

El trastorno por déficit de atención con hiperactividad (TDAH) es una alteración psicológica frecuente en niños, los síntomas que presenta son falta de atención, impulsividad e hiperactividad. Estas características hacen que los niños sean vulnerables a presentar otros desordenes y uno de ellos es el del lenguaje.

El interés por investigar los trastornos del lenguaje en niños con déficit de atención con hiperactividad surge a mediados de los 80's con la formulación del trastorno ya no solo como un trastorno de hiperactividad, esto desencadenó varias preguntas acerca del desarrollo del lenguaje en los niños. Se realizaron estudios que mostraron[8]:

1. Presentan una conducta lingüística un tanto peculiar.
2. Tienen más dificultades en la ejecución de ciertas tareas lingüísticas.
3. Realizan también una ejecución deficitaria de tareas que no requieren propiamente una respuesta verbal, pero en las el lenguaje actúa como mediador en la ejecución[8].

Los estudios de problemas de lenguaje en niños con déficit de atención se han abordado de la siguiente manera:

Como trastorno comórbido a trastornos psiquiátricos, mediante la observación de la íntima relación encontrada entre los trastornos del lenguaje y el TDA. Desde un punto de vista neuropsicológico, a través de la observación de la conducta lingüística de estos niños y a partir de ella deducir las posible disfunciones neurológicas. Desde un prisma cognitivo, con el seguimiento del modelo de procesamiento de la información. Y, en menor número de ocasiones, desde la visión de la patología del lenguaje.[8]

Sin embargo desde el aspecto funcional, los niños con TDAH no tienen muchos problemas en la utilización del lenguaje, su comunicación es eficaz, usando un lenguaje excesivo en situaciones comunes, pero si se les asigna una tarea lingüística especifica su rendimiento es malo. Un niño con TDAH es lingüísticamente funcional a menos que también presenten un trastorno específico del

lenguaje.

Los aspectos lingüísticos donde se observa que los niños con TDAH tienen especiales dificultades son los siguientes: Procesamiento fonológico y sintáctico, pero no en aspectos semánticos. Dificultades en tareas que requieren organización semántica. Tareas de memoria auditiva. Baker y Cantwell observaron, en la muestra de niños con TDAH que estudiaron, que el 78% de ellos presentaban problemas de articulación; el 69% problemas de procesamiento del lenguaje; dificultades en el lenguaje expresivo en un 58% y dificultades en lenguaje receptivo en el 34%. El índice de gravedad de estos trastornos era de moderado a leve en la mayoría de los niños y sólo grave en un 3%. Oram J et al. relacionan más los trastornos del lenguaje con el TDAH que los de habla. Los niños con TDAH también obtiene un rendimiento peor en los tests de fluencia verbal.[8]

Cabe resaltar que no todas las manifestaciones lingüísticas que pueden observarse en niños con TDAH son relevantes de la misma manera. Los problemas del habla, articulación y fluencia se relacionan en menor medida con el TDAH que los problemas de lenguaje.[9] Las características mas comunes de problemas de lenguaje por TDAH en niños son:

1. Presentan una conducta lingüística irregular con cierta ineficacia para ajustarse al contexto comunicativo y para comprender la intencionalidad comunicativa de su interlocutor.
2. Retraso en la adquisición de aspectos lingüísticos, en cuyo desarrollo influyen de forma decisiva los procesos atencionales tales como el código fonológico y el nivel morfológico, muy especialmente de comprensión y expresión de tiempos verbales.
3. Dificultades en la ejecución de tareas lingüísticas que demandan control inhibitorio como las tareas de fluidez léxica.
4. Obtienen peores resultados en tareas que requieren poner en marcha la capacidad de procesamiento simultáneo de la información como las tareas de procesamiento semántico y, especialmente, las que requieren pensamiento analógico lingüístico. Esta misma dificultad se observa en la ejecución de tareas de tipo metalingüístico, muy especialmente en las de conciencia fonológica.
5. También realizan una ejecución deficitaria en tareas que no exigen propiamente una respuesta verbal, en las que el lenguaje actúa como mediador en la ejecución.[9]

La relación entre el TDAH y los problemas de procesamiento del lenguaje conocidos como trastorno del procesamiento auditivo central (TPAC). El TPAC se define como el déficit en el procesamiento de señales auditivas que no puede ser adscrito a un déficit auditivo sensorial o periférico, o a un déficit intelectual. Este síndrome puede implicar la facilidad de distracción, problemas de memoria,

lectura y escritura.

3.3 Autismo

El autismo comúnmente se caracteriza por falta de capacidad de comunicación de calidad, así como de socialización, el problema con el desarrollo del habla ocurre en etapas tempranas del desarrollo de la comunicación. Los problemas motrices y de comunicación se observan desde niños, los cuales se reflejan en un problema de aprendizaje y de comportamiento.[10]

El estudio de Michele Noterdaeme et al. se han examinado a niños mediante un examen neurológico estándar y se han realizado pruebas como: funciones motrices delicadas como pintar, recortar, martillar, etc. y además funciones motrices rudimentarias tales como correr o saltar, entre otras pruebas han concluido que el desarrollo motriz relacionado con el trastorno específico del lenguaje mostró un menor desempeño en comparación a su grupo de control.

4 Modelos actuales

Los problemas de pronunciación afectan la presición y la calidad del habla y entendimiento del lenguaje. Para esto se han presentado dos modelos: el modelo de reestructuración léxica por Anne E. Fowler en 1991 y la teoría de distintividad fonológica por Elbro en 1998, dichos modelos muestran que los niños con escasa información de pronunciación de palabras se les dificulta distinguir a que letra pertenece el sonido y requieren una atención especializada para incrementar su percepción del habla. Los procesos de la percepción del habla incluyen un análisis auditivo preliminar, el análisis de las características de pronunciación y auditivas y la combinación de las características fonéticas en una representación fonológica.[6]

Un estudio llevado por Heather K.J. van der Lely, Stuart Rosen y Alastair McClelland en un niño de 10 años 3 meses que presenta trastorno específico del lenguaje, se creyó que su condición podría haber sido heredada de su padre y un tío paterno que también presentaban deterioro en sus habilidades con el lenguaje. Esto supone que pudo haber sido transmitido vía genética. Este niño presenta graves deterioros en su pronunciación y percepción del lenguaje, teniendo la habilidad de percepción de un niño de 5 años y 10 meses. Pero presenta un alto reconocimiento en palabras aisladas, es decir si se le presenta solo una palabra, el comprende de mejor manera su significado, asemejando a un niño de 7 años. La capacidad de expresarse verbalmente por medio de frases cortas es alta, de hecho es su manera de comunicación, pues en oraciones largas realiza pausas y omite artículos. Síntomas consistentes con TEL. Debido a este estudio se ha podido proporcionar evidencia de un déficit de lenguaje gramatical con implicaciones importantes en 2 áreas. La primera es que la posibilidad de que los déficit

auditivos y cognitivos que co-ocurren con el trastorno del lenguaje en algunos niños TEL pueden no ser la causa del deterioro. Y la segunda a existencia de una especialización determinada genéticamente de un subsistema en el cerebro necesaria para la gramática[4].

5. Conclusiones

El trastorno específico del lenguaje es una condición que puede afectar de manera severa el desarrollo del individuo que lo padece y debido a que este trastorno no se puede predecir aún como y cuando aparecerá, la manera de prevenir es casi imposible con los conocimientos actuales. El desorden conocido como dislexia ha sido definida e investigada desde finales del siglo XIX y junto con el déficit de atención con hiperactividad son los desordenes que mas presentan trastorno específico del habla y mucha de la investigación sobre este, es experimentado con individuos que presentan dichos desordenes. También se ha investigado el trastorno en personas que tienen daño cerebral o presentan autismo, pero se asume muchas veces que este trastorno es específicamente debido a daños en el cerebro y la forma en la que procesa la información con ese daño.

Se han propuesto algunos modelos por parte de neurólogos, psicólogos y psiquiatras para la atención adecuada de este y la correcta rehabilitación de las personas que sufren de este trastorno, ya se han identificado algunos muy importantes aspectos para diagnosticar este trastorno, los modelos propuestos indican un tratamiento lingüístico intensivo junto con técnicas de aprendizaje especializada. La investigación sobre este padecimiento no es escasa, sin embargo el análisis de los datos obtenidos y la propuesta de nuevos modelos o mejoras a los existentes avanza poco a poco, pues aun se encuentran muchas y variadas diferencias entre los individuos que presentan TEL, lo que dificulta crear un método estandarizado de atención.

References

1. Trastorno del lenguaje en niños. Biblioteca Nacional de Medicina de los EE.UU. Recuperado de https://www.nlm.nih.gov/medlineplus/spanish/ency/article/001545.htm el 18 de Febrero de 2016.
2. Johannes C. Ziegler, Catherine Pech-Georgel, Florence George, F.-Xavier Alario, Christian Lorenzi. 2005. Deficits in speech perception predict language learning impairment. Proceedings of the National Academy of Sciences of the United States of America.
3. H. K. J. Van Der. L Ely. L. S Tollwerck. 1996. A Grammatical Specific Language Impairment in Children: An Autosomal Dominant Inheritance?.Department of Psychology, Birkbeck College, University of London, United Kingdom.

4. Heather K.J. van der Lely, Stuart Rosen, Alastair McClelland. 1998. Department of Psychology, Birkbeck College, University of London, Malet Street, London. Department of Phonetics and Linguistics, and Department of Psychology, University College London. London, UK.
5. Rosario Ordz. Adelina Estévez. Mercedes Muñetón. 2014. El procesamiento temporal en la percepción del habla de los niños con dislexia.Servido de Publicaciones de la Universidad de Murcia. ISSN edición impresa: 0212-9728.Murcia, España.
6. Rosario Ortiz. Juan E. Jiménez. Mercedes Muñetón. Estefanía Rojas. Adelina Estévez. Remedios Guzmán. Cristina Rodríguez. Francisco Naranjo. 2008. Desarrollo de la percepción del habla en niños con dislexia. Universidad de La Laguna. Santa cruz de Tenerife, España.
7. R. Castro-Rebolledo. M. Giraldo-Prieto. L. Hincapié-Henao. F. Lopera. D.A. Pineda. 2004. Revista de neurologia. Grupo de Neurociencias. Universidad de Antioquia. Medellín, Colombia.
8. Ygual-Fernández. A. Miranda-Casas. J.F. Cervera-Mérida. 2000. Dificultades en las dimensiones de forma y contenido del lenguaje en los niños con trastornos por déficit de atención con hiperactividad. Revista de neurología clínica. acultad de Psicología. Universidad de Valencia. Valencia, España.
9. A. Miranda-Casas. A. Ygual-Fernández. F. Mulas-Delgado. B. Roselló-Miranda. R.M. Bó. 2002. Procesamiento fonológico en niños con trastorno por déficit de atención con hiperactividad: ¿es eficaz el metilfenidato?. Revista de neurología. Facultad de Ciencias de la Educación. Universidad de Valencia. Valencia, España.
10. Michele Noterdaeme. Katrin Mildenberger. Falk Minow. Hedwig Amorosa. 2002. Evaluation of neuromotor deficits in children with autism and children with a specific speech and language disorder. European Child & Adolescent Psychiatry. München, Alemania.

Capítulo 7

Análisis de selección de atributos con ganancia de información y X^2

Yuridiana Alemán, Darnes Vilariño, David Pinto
yuridiana.aleman@gmail.com, darnes@cs.buap.mx, and
dpinto@cs.buap.mx

Facultad de Ciencias de la Computación-BUAP
Av. San Claudio y 14 sur, CP: 72570, Puebla, Mexico
http://www.cs.buap.mx/

Resumen En este trabajo se realiza un análisis de la reducción de atributos para el proceso de clasificación supervisada. Para tal efecto, se utilizan dos conjuntos de *tweets* con diferente categoría (un conjunto clasifica género del autor y otro la polaridad del mensaje). Primeramente se clasifican con diversos conjuntos de categorías y posteriormente se realiza un análisis de ganancia de información y X^2 en los conjuntos de entrenamiento para obtener los atributos mas significativos. Para la creación de los modelos se utilizan los algoritmos de redes neuronales, máquinas de soporte vectorial (SMO) y bosque aleatorio, comparando la eficiencia de cada uno de ellos mediante el porcentaje de las instancias clasificadas correctamente. Los resultados obtenidos son mucho mejores utilizando menor cantidad de características, reduciendo así el tiempo de preprocesamiento y de generación de los modelos de clasificación.

Keywords: Ganancia de información, SMO, Bosque aleatorio, Selección de atributos, Redes neuronales, *Twitter*, X^2

1. Introducción

Actualmente, la información que circula por internet es inmensa, especialmente desde la masificación de la Web 2.0. Los volúmenes de datos se han incrementado en todas las áreas del conocimiento, toda esta información hace que las bases de datos cuenten con decenas e incluso cientos de miles de variables con un alto grado de información, la cual puede ser muy importante, pero también redundante o incluso irrelevante. Al momento de hacer un análisis de clasificación supervisada, la información redundante puede ocasionar el uso excesivo de tiempo y recurso computacional. En cambio, si se usan sólo los atributos mas significativos en en análisis de los datos, se obtendrán mejores resultados y sobre todo, se reducirán los tiempos de trabajo.

En este artículo se hace un análisis sobre la selección de atributos a traves del uso de la ganancia de información, es decir, que tan crucial es un atributo para determinar la clase principal. El documento está estructurado de la siguiente manera: En la sección 2 se mencionan algunos trabajos relacionados a la

selección de atributos aplicado a diferentes conjuntos de datos y algoritmos de clasificación, en la sección 3 se describe la metodología propuesta para el análisis. Posteriormente, en la sección 4 se discuten los resultados obtenidos usando todos los atributos obtenidos y aplicando los métodos de reducción de atributos, finalmente, la sección 5 resume las conclusiones obtenidas y el trabajo futuro.

2. Estado del arte

Se han realizado múltiples investigaciones sobre reducción de atributos, esto ha creado la existencia de varias técnicas, las cuales sólo son aplicables en determinadas circunstancias o bajo cierto tipo de conjutos de entrenamiento, entre los mas relevantes y relacionados a esta investigación se encuentran los siguientes:

En [1] se presenta un estudio sobre el funcionamiento de los métodos de selección de atributos. Por cada método de selección se generaron nuevos atributos que fueron probados con los algoritmos *Ant-Miner* [2] y C4.5. La comparación de los resultados se realizó en base a la exactitud y al número de reglas creadas. Los atributos seleccionados se aplican a la categorización de textos web. Los resultados ofrecen mejor precisión utilizando los atributos seleccionados.

En la investigación de [3] se desarrollan técnicas que permiten incorporar la selección de atributos para máquinas de soporte vectorial. La estrategia se basa en una eliminación secuencial hacia atrás y determina la contribución de cada atributo considerando aquel que impacta menos en el desempeño de la clasificación en un conjunto de validación independiente. Comenzando con todos los atributos disponibles, cada iteración eliminará aquellos atributos que afectan el desempeño predictivo hasta que se alcance un criterio de parada.

En [4] realizan reducción de atributos basados en n-gramas de palabras mediante un método de selección de características basada en reglas del texto multivariante denominado *Feature Relation Network* (FRN) que considera la información semántica y también aprovecha las relaciones sintácticas entre los n-gramas usados como atributos. El método FRN superó los resultados obtenidos con todos los n-gramas y el uso de otras características híbridas.

En cuanto al análisis de atributos basados en conteos, en [5] se utiliza una comparativa entre distintos conjuntos, pero se toma cada conjunto de características como un todo, sin analizar de manera individual cada atributo utilizado. En estos experimentos se muestran mejores resultados con el conteo de las categorías gramaticales utilizadas y la unión de todos los conjuntos analizados.

3. Metodología

La metodología propuesta se basa en la reducción del conjunto de características utilizadas en la clasificación para incrementar la exactitud, ésta se muestra en la figura 1.Para el análisis se utilizan dos conjuntos de *tweets* en inglés:

- El conjunto 1 fue extraído de la competencia *Lab on Uncovering Plagiarism, Authorship, and Social Software Misuse* (PAN 2013), el cual consiste en la

categorización por género y grupo de edad de blogs, socialmedia, *tweets* y foros. Para esta investigación se extrajo un porcentaje de los *tweets*, utilizando sólo la categoría correspondiente al género.

- El conjunto 2 corresponde a la competencia *International Workshop on Semantic Evaluation* (SemEval 2015, Tarea 2, subtarea B). Contiene *tweets* clasificados de acuerdo al sentimiento expresado (positivo, negativo y neutro).

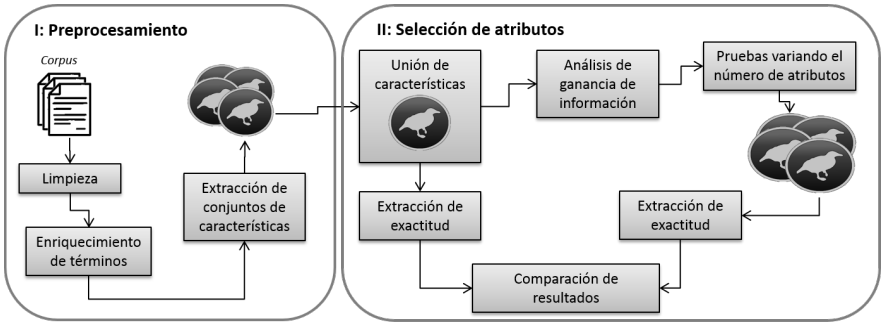

Figura 1. Metodología propuesta para la investigación

De ambos conjuntos sólo se cuenta con el conjunto de entrenamiento, pero se extrajo una parte de ellos para desempeñar el papel de conjunto de evaluación. La cantidad de *tweets* por conjunto y categoría se muestran en la tabla 1.Como puede observarse, a pesar de que los dos conjuntos son de *tweets*, contienen diferentes categorías, además, el primer conjunto está perfectamente balanceado y el conjunto de evaluación representa el 25 % del conjunto de entrenamiento, mientras que el segundo conjunto de datos , presenta 3 categorías, una de las cuales contiene muy pocas instancias respecto a las otras dos y no está balanceado.

Tabla 1. Conjuntos utilizados para el análisis

Conjunto 1			Conjunto 2		
Categoría	**Entrenamiento**	**Evaluación**	**Categoría**	**Entrenamiento**	**Evaluación**
Masculino	4,500	1,125	Positivo	2,336	352
Femenino	4,500	1,125	Negativo	864	176
			Neutro	3,099	422
TOTAL	**9,000**	**2,250**	**TOTAL**	**6,299**	**950**

El preprocesamiento se realiza implementando una expansión de los textos mediante el uso de algunos recursos léxicos (emoticones, palabras comunes en SMS y contracciones mas utilizadas). Una vez procesados los textos, se procedió

a la extracción de los siguientes conjuntos de características para obtener un total de 494 atributos:

1. **Conteos generales:** Se analizan características individuales extraídas durante la fase de preprocesamiento de los textos, tales como el uso de emoticones, contracciones, palabras mal escritas, palabras que inician con mayúscula, URLs, entre otras. Este conjunto abarca 13 características
2. **Palabras cerradas:** Dentro de esta categoría se clasifican grupos de palabras como preposiciones, conjunciones y determinantes. Cada palabra cerrada representa una característica, y el valor de ésta en cada instancia está dado por las veces que aparece en el texto. Se obtienen 194 palabras.
3. **Categorías gramaticales:** Se obtuvo la categoría gramatical de cada palabra dentro de los textos, para posteriormente realizar el conteo de de ellas. Para esto, se utilizó el *Stanford POS-tagger* [6], obteniendo 33 características.
4. **Signos:** Se contabilizan todos los signos de puntuación existentes (coma, punto y coma, interrogación, admiración, entre otros) obteniendo un conjunto de 32 características.
5. **Sufijos:** Se tomaron como características los sufijos existentes para el idioma inglés, al igual que en los conjuntos anteriores, cada sufijo representa una característica, y las veces que aparece en cada conversación es el valor para dicho atributo. En este conjunto se obtuvieron 122 características.
6. **Punto de transición:** Se experimenta con las palabras cercanas al punto de transición (PT), el cual sugiere que las palabras que caracterizan un texto no son ni las mas frecuentes ni las menos frecuentes, sino las que se encuentran en una frecuencia media de ocurrencia dentro del texto. La expresión para obtener el punto de transición se muestra en la ecuación 2, donde I_1 representa el número de palabras con frecuencia 1. Una vez obtenido este valor, las palabras con esa frecuencia se consideran las mas representativas del corpus y se utilizan como atributos en la clasificación. Esta técnica se utiliza en clasificaciones con muchos datos, en lugar de tomar en cuenta todo el vocabulario, se toman estas palabras ([7]). Se utilizaron las 100 palabras más próximas a este valor, sólo que en lugar de tomar la frecuencia, se utilizó el *tf-idf*, esta medida se conoce como la combinación de los conceptos de "frecuencia de términos" ($tf_{t,d} = Apariciones\ de\ t\ en\ d$) y "frecuencia inversa del documento" ($idf_t = log\frac{N}{df_t}$), la cual produce un peso para cada término en cada documento (fórmula 1).

$$tf - idf_{t,d} = tf_{t,d} \times idf_t \tag{1}$$

$$PT = \frac{\sqrt{1 + 8 * I_1} - 1}{2} \tag{2}$$

Tal y como se aprecia en la figura 1, ámbos conjuntos se clasificaron extrayendo las características anteriormente mencionadas, para ésto, se utilizaron los algoritmos de clasificación de máquinas de soporte vectorial (SMO) [8], Bosque

aleatorio [9] y redes neuronales[10], todos implementados en la herramienta *Weka* [11] y utilizando los parámetros que están por defecto.

En una segunda etapa de investigación, para reducir el conjunto de atributos utilizados en la clasificación y a la vez mejorar la exactitud obtenida, se utilizan los siguientes métodos individuales; cada uno con el método Ranker, que devuelve una lista ordenada de los atributos según su calidad:

1. **Ganancia de información:** Esta es una medida que evalúa el impacto que tiene el valor de un atributo respecto, y se obtiene mediante la ecuación 3

$$InfoGain(Clase, Atributo) = H(Clase) - H(Clase|Atributo) \quad (3)$$

2. **Ji cuadrada (X^2):** Calcula el valor estadístico Ji-cuadrado de cada atributo con respecto a la clase y así obtiene el nivel de correlación entre la clase y cada atributo.

4. Resultados obtenidos

Se crearon varios modelos de clasificación con los tres algoritmos utilizando todos los conjuntos de características, primero por separado y después unidos. Para evaluar la clasificación del conjunto de evaluación se analizó la exactitud, la cual se representa como el número de instancias bien clasificadas sobre el total en el conjunto. Como son varios conjuntos y se analizaron 3 algoritmos de clasificación, sólo se muestran los 5 mejores y los 5 peores resultados obtenidos. En la figura 2 se muestran resultados del conjunto 1 y en la figura 3 los resultados del conjunto 2. En cada gráfica se muestra el conjunto y el clasificador utilizado, en el caso del conjunto llamado "Todosçontiene todos los conjuntos (494 atributos), mientras que el conjunto llamado Ünion"son todos los atributos excepto el punto de trancisión (394 atributos).

En el conjunto 1, los 5 resultados menores se obtuvieron en su mayoría con redes neuronales, y se puede apreciar que con los 494 atributos, este algoritmo logra clasificar correctamente sólo el 39.3 % de los 2,250 *tweets* del conjunto de evaluación. El resultado mas elevado se obtiene con los mismos 494 atributos, pero utilizando como clasificador bosque aleatorio, llegando a un 70 % del total de instancias clasificadas correctamente.

En el conjunto 2 el nivel de exactitud es mucho mas bajo que en el conjunto 1, aunque al igual que en el primero, los resultados mas altos se dan con todos los conjuntos pero utilizano el algortmo SMO y los mas bajos son con redes neuronales. Analizando que el número de instancias y de clases por conjunto de evaluación, este conjunto tiene tres clases (positivo, negativo y neutro), por lo que las clasificaciones erroneas tienden a incrementarse, además, como se observa en la tabla 1, la clase de negativos está muy desbalanceada, al contrario del conjunto 1 que sólo tiene dos clases perfectamente balanceadas.

Para incrementar la exactitud obtenida, se ejecutan los métodos de reducción de atributos; en las figuras 4 y 5 se muestran los 20 atributos mejor ponderados con ganancia de información y X^2 para los conjuntos 1 y 2 respectivamente.

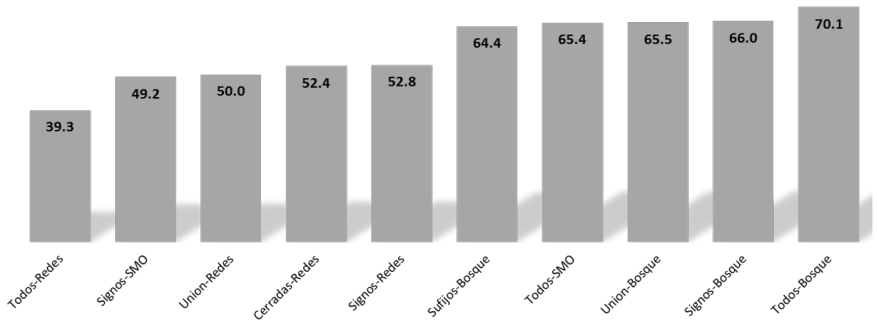

Figura 2. Porcentaje de elementos clasificados correctamente (5 mejores y 5 peores) en la clasificación del conjunto 1

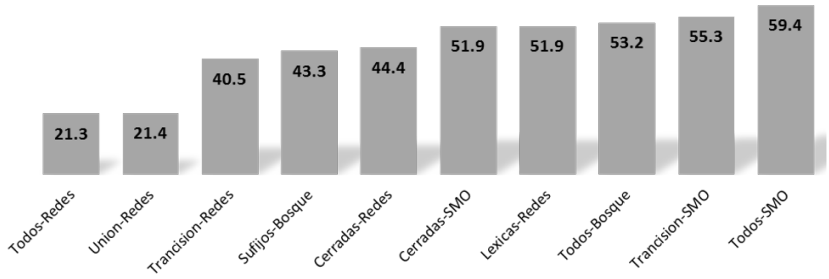

Figura 3. Porcentaje de elementos clasificados correctamente (5 mejores y 5 peores) en la clasificación del conjunto 2

Cada atributo está representado por el conjunto al que pertece (primeras letras mayúsculas) y el conteo que se realizó en su caso, por ejemplo el atributo "Sgato"pertenece al conjunto "Signosz representa el numero de # en los *tweets*, el atributo "TRANhe"pertenece a las palabras cercanas al punto de trancisión, específicamente a la palabra "hez CAVBP pertenece al conjunto de conteos de características gramaticales representando el número de verbos en 3ra persona tiempo presente (Representado por "VBP").

Como se puede observar, los dos conjuntos cuentan con diferentes atributos entre los que aportan mayor ganancia de información, predominando en el conjunto 1 los atributos correspondientes al punto de trancisión; aunque se dan

Ganancia de Información		Ji Cuadrada	
1 Sgato	11 TRANtmom	1 Sgato	11 TRANnptalk
2 TRANitem	12 TRANinstagr	2 TRANitem	12 TRANco
3 TRANskit	13 TRANco	3 TRANskit	13 TRANyahoolabs
4 TRANch	14 Sdiagonal	4 TRANch	14 TRANtmom
5 TRANpulse	15 TRANbeauty	5 TRANpulse	15 TRANinstagr
6 TRANmakeup	16 SguionBajo	6 URL	16 SUFectomy
7 URL	17 TRANtoo	7 CERRs	17 TRANtoo
8 TRANnptalk	18 SUFectomy	8 TRANmakeup	18 CERRtoo
9 TRANs	19 TRANhe	9 SguionBajo	19 TRANbeauty
10 TRANyahoolabs	20 TRANly	10 Sdiagonal	20 TRANly

Figura 4. Atributos con mayor ganancia de información para el conjunto 1

algunas diferencias en cuanto al método para la selección. Sin embargo, esta diferencia está dada por las características de los conjuntos de entrenamiento utilizados en cada caso, ya que el procedimiento es el mismo en ambos conjuntos.

Ganancia de Información		Ji Cuadrada	
1 Admiracion	11 Contracciones	1 Admiracion	11 Contracciones
2 CERRnot	12 TRANlove	2 TRANnot	12 TRANlove
3 CATPRP	13 CERRi	3 CATPRP	13 CERRi
4 CATRB	14 DosPuntos	4 CATRB	14 DosPuntos
5 Mayus	15 TRANwait	5 Mayus	15 TRANwait
6 URL	16 CERRat	6 URL	16 CERRit
7 TRANgood	17 CERRit	7 TRANgood	17 CERRwhy
8 Emoticones	18 Punto	8 Emoticones	18 Punto
9 TRANhappy	19 Numeros	9 TRANsmile	19 CERRat
10 TRANsmile	20 SUFiable	10 TRANhappy	20 SUFiable

Figura 5. Atributos con mayor ganancia de información para el conjunto 2

Una vez realizada la extracción de la ganancia de información y X^2 se ordenaron los atributos de mayor a menor valor y se realizaron pruebas con distinto número de ellos. Dados los bajos resultados de las redes neuronales en los primeros experimentos, sólo se utilizaron los algoritmos de bosque aleatorio y SMO. En la tabla 2 se muestra la exactitud obtenida por número de atributos y clasificador en cada experimento realizado utilizando los atributos generados por la selección con ganancia de información.

Se puede observar en la tabla que los resultados varían de acuerdo al algoritmo de clasificación y conjunto evaluado. Para el conjunto 1, con bosque aleatorio y 125 atributos se llegó a 73.11 % de exactitud (3 % más que utilizando los 405 atributos). Otro aspecto importante es que en todos los casos, la exactitud fué mas alta con bosque aleatorio, con resultados variando entre 69 % y 70 %, pero

Tabla 2. Exactitud obtenida en porcentaje y número de atributos utilizando la selección con ganancia de información

No. de	Conjunto 1		Conjunto 2	
atributos	**SMO**	**Bosque**	**SMO**	**Bosque**
25	50.356	68.089	58.316	56.632
50	58.222	69.422	59.158	58.737
75	61.067	70.622	59.790	58.842
100	62.489	72.000	59.790	57.79
125	60.622	72.044	60.737	59.263
150	61.333	70.978	60.842	58.947
175	62.000	71.556	59.579	58.211
200	62.489	69.511	59.263	58.737
225	61.556	71.600	59.053	58.842
250	63.289	71.511	60.526	56.947
275	62.533	70.667	60.211	58.947
300	62.800	71.067	60.526	57.579
325	63.067	70.622	59.053	56.737
350	63.422	69.778	59.263	56.421
375	63.556	71.378	59.263	57.053
400	63.600	70.711	60.526	57.053
425	63.956	69.467	60.000	56.000
450	64.044	70.133	59.368	57.474
475	64.311	70.000	59.474	57.263

obteniendo la exactitud más elevada entre los 125 y 200 atributos. Después de este punto, aunque es prácticamente imperceptible, los resultados empiezan a disminuir. Con SMO el resultado más alto fue con todos los atributos, llegano a un 64.3 %.

En el conjunto 2, con bosque aleatorio y 150 atributos se llegó a prácticamente 60 % de exactitud (7 % más que utilizando la totalidad de atributos). Al igual que en el conjunto 1, todos los experimentos obtuvieron mejores resultados con bosque aleatorio. Sin embargo, en este caso las exactitudes mas altas se presentan entre los 100 y los 175 atributos (a excepción de la clasificación con 325 atributos que obtiene 59 % de exactitud). Para este conjunto, el clasificador SMO obtuvo mejores resultados a partir de los 125 atributos llegando casi al 61 %.

De aquí se pueden obtener dos observaciones: Los resultados de bosque aleatorio pueden estar sesgados, ya que en su algoritmo, utiliza la ganancia de información para alegir la rama.ª seguir en la clasificación, y SMO es mas consistente cuando se trabaja con 3 categorías.

En la tabla 3 se muestran los resultados trabajando con los atributos obtenidos del análisis X^2. Para el conjunto uno, a pesar de que ya no está la influencia de la ganancia de información propia del bosque aleatorio, se observa que en la mayoría de los experimentos bosque aleatorio obtiene la mejor exactitud, superando el 73 % con 125 atributos (cabe observar que con 125 atributos también se logró la mejor exactitud en la tabla anterior), mientras que SMO llega a la ma-

Tabla 3. Exactitud obtenida en porcentaje y número de atributos utilizando la selección con X^2

No. de atributos	Conjunto 1		Conjunto 2	
	SMO	**Bosque**	**SMO**	**Bosque**
25	49.867	61.200	59.263	56.421
50	56.533	69.022	59.053	58.947
75	60.178	70.533	60.316	57.158
100	63.422	70.622	60.211	59.474
125	62.889	73.111	60.105	58.000
150	61.111	70.578	60.842	59.789
175	60.533	70.844	60.211	58.000
200	60.444	71.289	60.316	57.368
225	61.111	70.711	59.79	56.947
250	61.022	70.044	59.158	56.632
275	61.911	69.733	60.526	57.263
300	62.978	70.000	60.526	57.684
325	62.800	69.733	60.211	59.158
350	62.622	70.267	59.263	56.737
375	62.533	69.867	59.158	57.263
400	63.689	69.778	60.211	55.368
425	63.556	70.711	59.790	56.211
450	63.289	69.822	60.737	57.263
475	63.156	69.644	59.684	56.211

yor exactitud a medida que se acerca al total de atributos utilizados, alcanzando un 63.3 %.

En el conjunto 2, al igual que utilizando ganancia de información, SMO obtiene mejores porcentajes de bosque aleatorio, llegando a 60.84 % con 150 atributos; estos resultados varían muy poco de los obtenidos anteriormente. Con el algoritmo de bosque aleatorio se alcanza un 59 % utilizando de 100 a 175 atributos, lo cual es un poco mejor que utilizando ganancia de información.

5. Conclusiones y trabajo futuro

De los experimentos reportados en esta investigación se puede concluir lo siguiente:

- Los resultados de los dos conjuntos analizados difieren mucho, ya que, aunque en ámbos conjuntos se manejan *tweets*, el conjunto 1 está balanceado y tiene sólo dos clases, mientras que el conjunto 2 presenta 3 clases, una de las cuales tiene muy pocas instancias para su predicción. Esto mismo hace que los resultados del conjunto con dos clases sean mucho mayores que los del conjunto con tres clases.
- Al reducir los atributos utilizados, se muestra que utilizando de 100 a 175 atributos, la exactitud de la clasificación de incrementan hasta en 3 % en el caso del conjunto 2 y 1 % para el conjunto 1. Aunque no es mucho el

incremente, el utiizar menos atibutos permite reducir el tiempo de creación de los modelos, así como el recurso computacional utilizado.
- Con el algoritmo de bosque aleatorio se obtiene la mejor exactitud para la clasificación binaria, mientras que SMO da mejores resultados cuando se trabaja con tres categorías

Como trabajo futuro, se tiene planeado analizar este mismo proceso de selección de atributos utilizando conjuntos de textos de otra naturaleza, a fin de determinar si los resultados son igual de factibles o sólo se incrementa la exactitud en textos de redes sociales. Además, analizar otros procedimientos de reducción de atributos para seguir incrementando los resultados. Otra área importante de investigación se centra en el análisis de los atributos seleccionados en cada una de las clases por separado.

Referencias

1. Saian, R., Ku-Mahamud, K.: Comparison of attribute selection methods for web texts categorization. In: Computer and Network Technology (ICCNT), 2010 Second International Conference on. (2010) 115–118
2. Parpinelli, R.S., Lopes, H.S., Freitas, A.A.: Data mining with an ant colony optimization algorithm. IEEE Transactions on Evolutionary Computation **6** (2002) 321–332
3. Maldonado, S., Pérez, J., Weber, R., Labbé, M.: Feature selection for support vector machines via mixed integer linear programming. Information Sciences **279** (2014) 163 – 175
4. Abbasi, A., France, S., Zhang, Z., Chen, H.: Selecting attributes for sentiment classification using feature relation networks. IEEE Transactions on Knowledge and Data Engineering **23** (2011) 447–462
5. Alemán, Y., Vilariño, D., Pinto, D., Tovar, M.: Detección de depredadores sexuales utilizando un sistema de consulta y clasificación. Research in Computing Science **72** (2014) 9–21
6. Toutanova, K., Klein, D., Manning, C., Singer, Y.: Feature-rich part-of-speech tagging with a cyclic dependency network. In: Human Language Technology Conference (HLT-NAACL 2003). (2003)
7. Jiménez-Salazar, H., Pinto, D., Rosso, P.: Uso del punto de transición en la selección de términos índice para agrupamiento de textos cortos. Procesamiento del Lenguaje Natural **35** (2005)
8. Platt, J.C.: Advances in kernel methods. MIT Press, Cambridge, MA, USA (1999) 185–208
9. Breiman, L.: Random forests. Mach. Learn. **45** (2001) 5–32
10. Mitchell, T.M.: Machine Learning. 1 edn. McGraw-Hill, Inc., New York, NY, USA (1997)
11. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. SIGKDD Explor. Newsl. **11** (2009) 10–18

Parte II

Aplicaciones en la Ingeniería del Lenguaje y del Conocimiento

Capítulo 8

Aplicación móvil para recuperación de información usando preferencias del usuario

Ana B. Rios-Alvarado, Edgar Tello-Leal, Alan Díaz-Manríquez, Tania Y. Guerrero-Meléndez

Facultad de Ingeniería y Ciencias - Posgrado e Investigación
Universidad Autónoma de Tamaulipas
{arios, etello, amanriquez, tyguerre}@uat.edu.mx

Resumen Actualmente, los dispositivos móviles se han convertido en herramientas indispensables en las actividades diarias de una gran parte de la población. De manera especial, se ha incrementado el interés por el acceso a la información de forma precisa y rápida, lo que hace necesario contar con nuevas aplicaciones para incorporar características adicionales a la búsqueda de información a través de un dispositivo móvil. En este artículo, se presenta el diseño de una aplicación móvil que integra un modelo de representación semántico de las preferencias de usuario en una aplicación de software para la búsqueda de información para dispositivos móviles. La integración de tal esquema en la personalización de la consulta, así como en la presentación de los resultados permitirá recuperar datos precisos y útiles de acuerdo al usuario y el dispositivo móvil.

Keywords: Recuperación de información, dispositivo móvil, ontologías

1. Introducción

El uso de los dispositivos móviles comprende aplicaciones diversas en los campos de entretenimiento, ofimática, redes sociales, monitoreo vial, monitoreo médico (por ejemplo: para pacientes con padecimientos como diabetes o hipertensión), búsqueda de información, entre otras. Sin embargo, una de las acciones más recurrentes de los usuarios de dispositivos móviles es la búsqueda de información, los usuarios requieren de información precisa de acuerdo a sus necesidades y preferencias; por consiguiente, un dispositivo móvil se comporta como una herramienta de acceso a la información. Se han propuesto aplicaciones de recuperación de información para dispositivos móviles en áreas muy específicas como biomédica [9] y diseño de modas [3]. Otros trabajos han incorporado ontologías de dominio específico como geo-localización [8], búsqueda de aplicaciones móviles [6]. Algunos han considerado crear un contexto a partir de la ubicación del usuario [2]. También existen trabajos [13], [1] donde se ha considerado aprovechar los componentes de un dispositivo móvil (sensores) para obtener información del contexto del usuario. Sin embargo, aún con el amplio desarrollo de aplicaciones

para dispositivos móviles, la definición e integración de un esquema que considere las preferencias (profesión, hobbies, aficiones culturales o deportivas, edad, ubicación actual, entre otras) del usuario, independientemente de la aplicación, ha sido poco estudiada.

La personalización de las consultas ha sido abordado en la literatura considerando la interacción del usuario con un conjunto de conceptos y sus relaciones. Este conjunto de elementos se ha organizado en una estructura llamada ontología de perfil de usuario [10]. La definición de ontología se ha ido enriqueciendo en los últimos años, debido al gran auge que ha adquirido en el desarrollo de nuevas aplicaciones, como por ejemplo, en la búsqueda de información, en donde se pueden obtener resultados pertinentes. Noy y McGuinness [11] definen una ontología como una especificación formal y explícita de conceptos en un dominio. Una ontología define los términos a utilizar para describir y representar un área de conocimiento. Las ontologías son utilizadas por las personas o por las aplicaciones de software que necesiten compartir un área temática específica.

Por lo tanto, se propone diseñar e implementar una ontología para la representación de preferencias del usuario e integrarlo a una aplicación de recuperación de información para un dispositivo móvil, con el propósito de obtener resultados precisos a las consultas de información. En particular, la aplicación móvil tiene como objetivos permitir el registro de las preferencias del usuario y generar una instancia del perfil del usuario con el fin de integrar información adicional a una consulta y así obtener resultados relevantes para el usuario. La consulta se extenderá automáticamente y se enviará a un buscador web, los resultados recuperados se mostrarán ordenados tomando en cuenta las preferencias del usuario.

A continuación, en la Sección 2 se presenta una revisión de las propuestas relacionadas con recuperación de información en dispositivos móviles. Posteriormente, en la Sección 3 se incluye el planteamiento del problema y se describe la propuesta. En la Sección 4 se presenta la metodología para el desarollo de la ontología que represente el perfil del usuario. En la Sección 5 se describen algunos componentes principales de la aplicación y los experimentos realizados. Finalmente en la Sección 6 se dan las conclusiones.

2. Trabajos relacionados

Algunos de los trabajos que han abordado el tema de recuperación de información en dispositivos móviles se concentran en crear aplicaciones para manipular información en un dominio específico. Millán *et al.*, [9] presentan un Sistema de Recuperación de Información en dispositivos móviles en el dominio biomédico donde proponen un algoritmo de agrupamiento llamado *Clustering on Mobile Medical Devices* (CLUMMED), el cual hace un agrupamiento posterior a la recuperación. CLUMMED integra documentos del ámbito biomédico y organiza los resultados de búsqueda en carpetas que almacenan los documentos que están semánticamente relacionados, esto facilita la navegación entre ellos. DRESSER, presentada por Buitrago [3], es una aplicación web personalizada

para consultar y compartir información en el contexto de la moda de acuerdo a las preferencias (edad, colores preferidos, religión, género, etc.) y el contexto (época del año) del usuario. Por su parte, Xia *et al.* [13] proponen el uso de una ontología para ubicación de puntos de interés y al mismo tiempo recuperar puntos de interés relacionados semánticamente. La aplicación desarrollada por Datta *et al.* [6] integra capacidades semánticas en la búsqueda de aplicaciones móviles, además incorpora un mecanismo de puntuación para los resultados con base en el usuario y el grupo de desarrollo de la aplicación. Biancalana *et al.* [2] presentan un enfoque para la recomendación de puntos de interés, por ejemplo, restaurantes o eventos culturales, tomando en cuenta la ubicación actual del usuario y recomendaciones de lugares a partir de las redes sociales o páginas web.

Otros trabajos presentan el uso de sensores u otros componentes integrados para obtener información del contexto del usuario. Kang *et al.* [8] presentan el desarrollo de un sistema de recuperación de información regional para dispositivos móviles denominado *Mobile SeoulSearch*, el cual tiene como objetivo mejorar las búsquedas de información regional y superar las limitaciones de los dispositivos móviles, como el tamaño de la pantalla. Este sistema proporciona información relacionada con la posición específica en la que el usuario se encuentra en tiempo real, de manera que se aprovecha la movilidad de los dispositivos. El sistema *JIT MobIR* (del inglés Recuperación de Información en Tiempo Real) [1], utiliza los sensores incorporados en el dispositivo móvil, con el objetivo de identificar el contexto del usuario. Si se utiliza el contexto de la información del usuario, se puede evitar la recuperación de información que no es relevante para el usuario, haciendo eficiente la búsqueda en el dispositivo móvil.

3. Descripción de la propuesta

Los dispositivos móviles inteligentes (*smartphones o tablets*) permiten el acceso desde buscadores tradicionales, estos recuperan una gran cantidad de enlaces, de los cuales sólo una pequeña parte es realmente útil para el usuario. Por otro lado, se tiene el acceso a información a través de redes sociales, donde la consulta de información está limitada a estar registrado en esa red social y se recuperan las entidades con las que se tenga una asociación. También debe considerarse que la consulta realizada utilizando palabras clave, da espacio a consultas imprecisas o ambiguas. Una de las técnicas más conocidas para lidiar con estos problemas es la expansión de términos de la consulta, que ha sido adoptada por sistemas comerciales, especialmente para sistemas de escritorio o redes locales. La principal ventaja de agregar términos a la consulta es que se puede ampliar la cobertura en el espacio de búsqueda [4]. Por ejemplo, si tenemos la consulta «UAT» y agregamos la ubicación «Ciudad Victoria» podemos recuperar con una mejor posición la página de la «Universidad Autónoma de Tamaulipas» en lugar de recuperar «Universidad Autónoma de Tlaxcala».

Por otro lado, cada usuario tiene características y preferencias distintas a considerar en el momento de llevar a cabo la búsqueda de información. La bús-

queda web personalizada pretende ayudar a delimitar el contexto del usuario a partir del cual se puede plantear un conjunto de características que facilite el acceso al conocimiento. El contexto del usuario puede ser determinado a partir de la consulta, también llamada necesidad de información, la semántica de la consulta y el conjunto de intereses que describen al usuario [12].

Actualmente, las aplicaciones de búsqueda en dispositivos móviles no considera la integración de un perfil de usuario y un servicio de búsqueda con una consulta personalizada automáticamente. Además es importante mostrar los resultados de manera eficiente para el dispositivo móvil. Por lo tanto, dado que un dispositivo móvil es personal, se puede obtener información adicional a partir de sus componentes de hardware y almacenar información semántica de forma local, entonces ¿será posible diseñar y desarrollar una aplicación que mediante el uso de un esquema semántico para representar las preferencias del usuario y con los resultados obtenidos presente un listado de acuerdo al usuario de tal dispositivo?

Se propone desarrollar una aplicación para dispositivos móviles que haga uso de una representación semántica de las preferencias del usuario y así mejorar la precisión en la recuperación de información. De manera particular, se pretende definir una ontología para describir las preferencias del usuario, definir e implementar un método para la personalización de la consulta que considere el uso las preferencias del usuario e integrar todo en una aplicación para dispositivo móvil.

Figura 1. Fase de consulta

Figura 2. Fase de recuperación

Las fases principales que se llevarán a cabo son la consulta y la recuperación de información. En la Figura 1 se muestra el proceso de la fase de consulta donde el usuario realizará una consulta desde el dispositivo móvil. La consulta será personalizada usando el perfil del usuario y posteriormente será enviada a un servicio de búsqueda web. Para la fase de recuperación, mostrada en la Figura 2, los resultados obtenidos se ordenarán considerando las preferencias del usuario y finalmente serán mostradas en una lista comenzando con aquellos que son más precisos para el usuario.

4. Ontología de perfil de usuario

En esta sección se presentan las etapas que se siguieron para la definición y construcción de la ontología que soporta el conjunto de preferencias del usuario. Este esquema semántico es almacenado en el dispositivo móvil y es utilizado por la aplicación de búsqueda al momento de extender la consulta.

En los últimos años han surgido una gran cantidad de metodologías, lenguajes y herramientas que permiten la construcción de una ontología. Corcho [5] hace un análisis de metodologías, herramientas y lenguajes para la construcción de una ontología y propone que para su construcción se deben considerar tres aspectos: ¿qué metodología usar para la construcción de una ontología?, ¿qué herramienta de apoyo usar para el desarrollo de una ontología? y ¿qué lenguaje usar para la implementación de una ontología?. La comparación entre las metodologías analizadas se basa en el grado de dependencia que existe entre la ontología y

la aplicación final en donde se le utiliza. En este trabajo se ha considerado la metodología propuesta por Noy y McGuinness [11] con las siguientes etapas:

1. *Determinar el dominio y alcance de la ontología.* En esta etapa se estableció que el dominio de la ontología es *perfil de usuario*. Su utilidad radica en permitir representar las características y preferencias principales de un usuario. Donde el usuario es el interesado en consultar y buscar información a través de un dispositivo móvil.
2. *Considerar el uso de ontologías o recursos existentes.* Tomando en cuenta a Golemati *et al.* [7] se tiene que el conjunto de clases de alto nivel necesario para construir un esquema del perfil de usuario son: *Person*, *Characteristic*, *Ability*, *LivingConditions*, *Contact*, *Preference*, *Interest*, *Activity*, *Education* y *Profession*. Algunas clases pueden especializarse de acuerdo al área de expertise del usuario o el contexto dónde el usuario usará la aplicación. Se han considerado las clases *Person*, *Preference* y *Profession* para integrarlas a la ontología desarrollada en este trabajo.
3. *Hacer una lista de los términos importantes.* Para obtener la lista de términos importantes y definir las clases se realizó un análisis de los formularios que deben llenarse para el registro de un nuevo usuario en distintas redes sociales (por ejemplo: Facebook, Twitter, Instagram, LinkedIn, entre otras). Se obtuvo una lista de las características que definen a un usuario y se clasificaron por tipo de dato que describen.
4. *Definir las clases y las relaciones jerárquicas.* Una vez que se obtuvieron las características principales de un usuario, se definieron las siguientes clases principales: *Person*, *Occupation*, *Place*, *Picture*, *Organization*, *Profession*, *Interests* y *Preferences*. Se estableció que la clase *Interests* y *Preferences* son equivalentes. En total se identificaron 120 clases para representar el perfil de usuario. A continuación, se realizó un análisis para establecer relaciones jerárquicas. La mayoría de las relaciones jerárquicas recaen sobre la clase *Preferences*. Por ejemplo, se tiene que *FilmGenre*, *MusicGenre*, *Bussiness*, *Culture*, *Education*, *Entertainment*, *FoodAndDrinks*, *LifeStyle*, *Parenting*, *Politics*, *ScienceAndTechnology*, *Sports* and *Travel* son subclases de *Preferences*. Asimismo, se estableció que las clases *Action*, *Adventure*, *Comedy*, *CrimeAndGanster*, *Drama*, *EpicsHistorical*, *Horror*, *Musicals*, *ScienceFiction*, *War* y *Westerns* son subclases de *FilmGenre*. Por otro lado, por ejemplo, las clases *Employee* y *Student* son subclases de *Occupation*. Se identificaron 125 relaciones jerárquicas.
5. *Definir las propiedades de las clases.* Posteriormente, se realizó la identificación de relaciones entre clases, se obtuvieron 20 propiedades de objetos, por ejemplo, *<User,worksAt,Company>*, *<User, livesAt,City>*, *<User,hasOccupation,Occupation>*, *<User, prefersTo, Preferences>*, entre otras. También se definieron algunas propiedades de datos, por ejemplo, *<User,hasName,string>*, *<User,hasUserName,string>*, *<User,hasPassword, string>*, *<User,hasBirthDate,dateTime>*. En total se tienen 18 propiedades de datos.
6. *Crear la instancia.* Una instancia de perfil de usuario es creada para cada usuario en su dispositivo móvil, esto se hace a través de la aplicación móvil.

Figura 3. Jerarquía de clases en Protégé

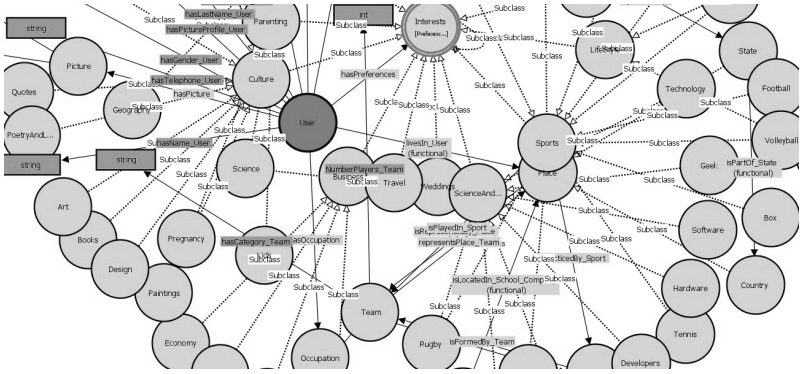

Figura 4. Fragmento de las clases de la ontología de perfil de usuario

Como herramienta de apoyo se usó Protégé[1] para realizar la edición de la ontología y el lenguage OWL[2] para implementarla. En la Figura 3 se puede observar un fragmento de la jerarquía de clases implementada usando Protégé. En la Figura 4 se muestra de manera gráfica un fragmento de las clases implementadas y sus propiedades.

[1] http://protege.stanford.edu/
[2] https://www.w3.org/TR/owl-ref/

5. Aplicación móvil para consulta y recuperación de la información

A continuación, se describe el funcionamiento principal de la aplicación móvil para consulta y recuperación de la información. Cuando un usuario requiere realizar una consulta usando la aplicación propuesta, primero deberá realizar un registro, en el cual el usuario introduce sus datos generales y llena un formulario con datos de sus preferencias, además introduce un nombre de usuario y contraseña. Al registrarse la aplicación móvil generará una instancia del perfil de usuario. El usuario puede decidir si ingresa una consulta o sale de la aplicación. Si el usuario decide realizar una consulta, la aplicación móvil se encargará de agregar parámetros para personalizar la consulta. Posteriormente, se enviará la consulta personalizada al buscador web. Los resultados recuperados se filtran para mostrar los enlaces devueltos por el buscador y la aplicación móvil los muestra ordenados de acuerdo a las preferencias.

Las medidas tradicionalmente empleadas para medir la relevancia de la recuperación de información son la *precisión*, *exhaustividad* y *promedio de la efectividad*. En este caso, para la evaluación de la aplicación se ha empleado la medida de precisión. La precisión mide el porcentaje de documentos recuperados que resultan relevantes con el tema de la pregunta y se calcula dividiendo el total de documentos relevantes recuperados entre el total de documentos recuperados. Cabe mencionar que la exhaustividad no se reporta debido al alcance de documentos que se tiene usando el buscador web.

Para evaluar la precisión de la aplicación propuesta se definieron 10 perfiles de usuario distintos y se plantearon 100 consultas, las cuales se ejecutaron de forma tradicional mediante la aplicación de búsqueda web de Google, así como también se ejecutaron las mismas consultas considerando la *consulta extendida* con datos del perfil. Entonces, se midió la precisión tomando en cuenta los enlaces devueltos como resultado de la consulta y cada usuario manualmente determinó cuántos eran relavantes a la consulta dada sus condiciones actuales. Por ejemplo, la Tabla 1 muestra los resultados recuperados de una consulta simple y una consulta extendida. Considerese que un usuario recientemente llega a *Ciudad Victoria* y su género de cine favorito es el *Drama* y desea consultar los cines cercanos. Dadas las características del dispositivo móvil quizá solo ejecute una consulta con la palabra *cinema* y por tanto obtendra pobres resultados. La consulta extendida en forma automática agregaría las palabras *Drama* y *Ciudad Victoria* a la consulta y como se puede observar en la Tabla 1 se recuperan enlaces relevantes a la consulta del usuario.

La Tabla 2 muestra el promedio de la precisión obtenido para el conjunto de consultas evaluado por 10 usuarios distintos. Cabe destacar que en algunos casos la precisión es baja debido a las consultas que pueden ser muy especializadas, es decir, en las que los datos de las preferencias no son representativas de la información que el usuario desea. Por otro lado, la precisión también puede ser baja si en un determinado momento el usuario cambia de interesés y el perfil conserva datos iniciales.

Consulta: *cinema*	
Resultados	Relevante
https://cinemex.com/cinema	No
https://www.youtube.com/watch...	No
https://es.wikipedia.org/wiki/Cinema	No
https://en.wikipedia.org/wiki/Cinema	No
https://es.wikipedia.org/wiki/Categor...	No
https://www.facebook.com/alfhaville/	No
http://www.cairocinemacafe.com/	No
http://www.funkycinema.com/	No
https://www.blackmagicdesign.com/mx/...	No
Consulta extendida: *cinema Drama Ciudad Victoria*	
Resultados	Relevante
http://www.cinesdemexico.com/ ... /cinemex-victoria	Si
http://www.cinesdemexico.com/ ... /cinepolis-plaza-campestre	Si
http://www.cinepolis.com/cartelera/cd-victoria	Si
https://cinemex.com/cine/200/cinemex-cd-victoria	Si
http://www.cartelerasdecine ... cinepolis-plaza-campestre/	Si
http://www.cartelerasdecine. ... cinemex-mmc-ciudad-victoria /	Si
http://www.cinefis.com.mx/cinepolis-plaza-campestre/ ...	Si
http://www.cinefis.com.mx/cinemex-victoria/cine/8715	Si
https://twitter.com/cineteca_cct	Si
http://www.sensacine.com/cines/cine/E0632/	No

Tabla 1. Ejemplo de consulta y resultados recuperados

Perfil de Usuario	Precisión (%)
Usuario 1	75.65
Usuario 2	63.21
Usuario 3	78.36
Usuario 4	54.32
Usuario 5	87.36
Usuario 6	45.32
Usuario 7	76.32
Usuario 8	78.36
Usuario 9	68.32
Usuario 10	78.36

Tabla 2. Promedio de precisión calculada por perfil de usuario

A continuación, como parte del desarrollo de la aplicación se muestran detalles de la implementación de las interfaces gráficas de usuario. La Figura 5 muestra la interfaz de usuario para el proceso de registrar el perfil de usuario. El usuario deberá introducir datos como nombre, e-mail, fecha de nacimiento, género, ocupación, área de interés (ciencia, deportes, política, tecnología, moda y cultura). Además puede ingresar comida favorita, deporte favorito y partido político predilecto. De manera opcional se puede activar el GPS para guardar datos sobre la ubicación dentro del esquema.

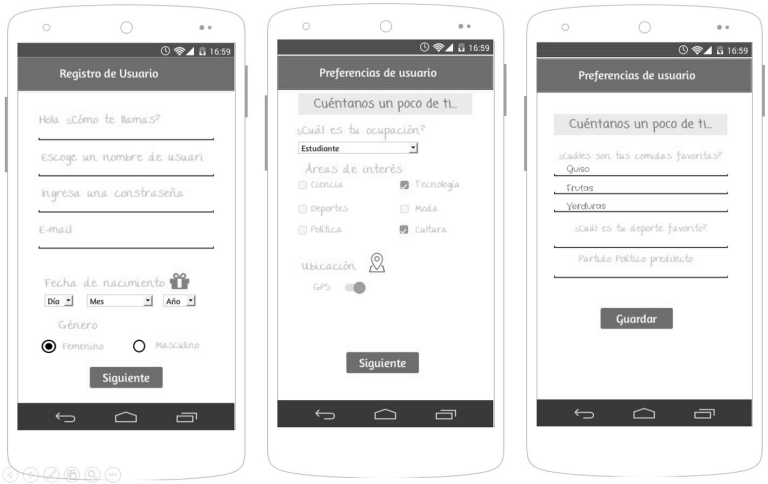

Figura 5. Interfaz gráfica de usuario: registrar perfil

La Figura 6 muestra la interfaz de usuario para introducir una consulta y la interfaz que muestra los resultados recuperados. El usuario deberá escribir su consulta y tendrá la opción de seleccionar el idioma y el tipo de contenido sobre el cual desea hacer la búsqueda. La interfaz de resultados despliega una lista de los recursos recuperados a partir de un buscador web.

6. Conclusiones y trabajo futuro

Se ha abordado la recuperación de información haciendo uso de una ontología para representar el perfil del usuario. Tal enfoque ha sido implementado en una aplicación móvil. Las características del dispositivo móvil proveen un buen escenario sobre el cual incorporar una instancia de perfil de usuario, dado que el dispositivo móvil es de uso personal.

Con esta aplicación se pretende satisfacer las necesidades de información de forma eficiente y con mayor precisión de acuerdo a características de ocupación, interesés y ubicación sin pertenecer a una red social. Además, la creación de la

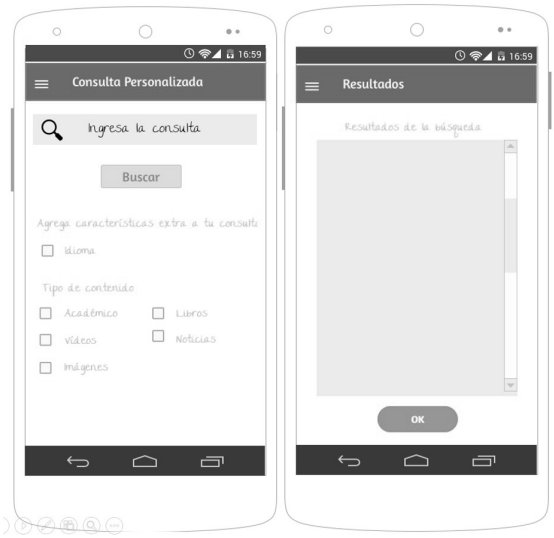

Figura 6. Interfaz gráfica de usuario: realizar consulta y mostrar resultados.

instancia del perfil de usuario puede ser útil a otras aplicaciones del dispositivo móvil, por ejemplo, para la sugerencia de descarga o eliminación de aplicaciones para el dispositivo móvil, también para el registro de información médica y la clasificación de mensajes recibidos. Como trabajo futuro se contempla la evaluación del proceso de recuperación de información considerando distintas instancias para el perfil del usuario.

Referencias

1. Alidin, A., Crestani, F.: Context acquisition in just-in-time mobile information retrieval. In: International Conference on Information Retrieval Knowledge Management (CAMP) 2012. pp. 203–207 (2012)
2. Biancalana, C., Gasparetti, F., Micarelli, A., Sansonetti, G.: An approach to social recommendation for context-aware mobile services. ACM Transanctions on Intelligent Systems and Technology 4(1), 10:1–10:31 (2013)
3. Buitrago Herrera, S.: Aplicación web personalizada para consultar y compartir Información en el contexto de la industria de la moda. Master's thesis, Pontificia Universidad Javeriana, Bogota, Colombia (2014)
4. Carpineto, C., Romano, G.: A survey of automatic query expansion in information retrieval. ACM Computing Survey 44(1), 1:1–1:50 (2012)
5. Corcho, O., Fernández-López, M., Gómez-Pérez, A.: Methodologies, tools and languages for building ontologies. where is their meeting point? Data & knowledge engineering 46(1), 41–64 (2003)
6. Datta, A., Dutta, K., Kajanan, S., Pervin, N.: Mobilewalla: A mobile application search engine. In: Zhang, J., Wilkiewicz, J., Nahapetian, A. (eds.) Mobile

Computing, Applications, and Services, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, vol. 95, pp. 172–187. Springer Berlin Heidelberg (2012)
7. Golemati, M., Katifori, A., Vassilakis, C., Lepouras, G., Halatsis, C.: Creating an ontology for the user profile: Method and applications. In: Proceedings of the first RCIS conference. pp. 407–412 (2007)
8. Kang, Y., jin Kwon, Y.: Mobile seoulsearch: Automatic mobile regional information retrieval system based on web. In: 21st International Conference on Data Engineering Workshops, 2005. pp. 1245–1245 (2005)
9. Millán, M., Muñoz, A., de la Villa, M., Maña, M.J.: A biomedical information retrieval system based on clustering for mobile devices. Procesamiento del Lenguaje Natural 45, 255–258 (2010)
10. Mohammed, N., Duong, T., Jo, G.: Contextual information search based on ontological user profile. In: Pan, J.S., Chen, S.M., Nguyen, N. (eds.) Computational Collective Intelligence. Technologies and Applications, Lecture Notes in Computer Science, vol. 6422, pp. 490–500. Springer Berlin Heidelberg (2010)
11. Noy, N.F., Mcguinness, D.L.: Ontology development 101: A guide to creating your first ontology. Tech. rep., Stanford University (2001)
12. Sieg, A., Mobasher, B., Burke, R.: Ontological user profiles for personalized web search. In: Proceedings of the 5th Workshop on Intelligent Techniques for Web Personalization, Vancouver, Canada. pp. 84–91 (2007)
13. Xia, Y., Luo, S., Zhang, X., Bae, H.Y.: A novel retrieval method for multimodal point of interest data. International Journal of Multimedia & Ubiquitous Engineering 9(7) (2014)

Capítulo 9

Aplicación Web para la Evaluación de Ontologías de Dominio

Karen Vazquez[1] and Mireya Tovar[1]

[1]Facultad de Ciencias de la Computación
Benemérita Universidad Autónoma de Puebla
Ciudad Universitaria, 72570 Puebla, Mexico
{krnlet,mireyatovar}@gmail.com

Abstract. El trabajo realizado consiste en el desarrollo de una aplicación web específicamente para la Evaluación Automática de Ontologías de Dominio Restringido. Se diseñó con apoyo de la metodología: Método de Diseño Hipermedia Orientado a Objetos. La Aplicación web se encuentra separada por usuarios comunes y usuarios expertos, los segundos centrado su especialidad de dominio. El sistema también incluye los suficientes apartados para que los usuarios tengan un eficiente resultado de acuerdo a la ontología a evaluar; desde Preprocesamiento, Descubrimiento Automático Y Evaluación cada uno con los recursos necesarios. El resultado del sistema es una interfaz practica y cómoda, además de contar con un diseño agradable a la vista.

Keywords: Aplicación, web, ontologias, OOHDM, evaluación

1 Introducción: Consultas de Evaluación en la web

Una aplicación web es un sitio web que contiene páginas con contendió sin determinar, parcialmente o en su totalidad, reciben este nombre porque se ejecutan en internet, es decir que los datos o los archivos en los que trabajas son procesado y almacenados dentro de la web, por lo que en general no necesitan ser instaladas en un ordenador.[1]

"En cualquier momento, lugar y desde cualquier dispositivo podemos acceder a este servicio, solo necesitamos una conexión a internet y nuestros datos de acceso, que por lo general son el nombre de usuario y contraseña"

Las ontologías son actualmente materia de investigación, desarrollo, y aplicación en disciplinas relacionadas con la computación, la información y el conocimiento. Los sistemas de información (SI) son esencialmente artefactos de conocimiento que capturan y representan el conocimiento sobre ciertos dominios.[2]

Las ontologías de dominio son un sistema de representación del conocimiento que se pueden organizar en estructuras taxonómicas y ontológicas de conceptos de algún área o dominio de conocimiento específico [3]. Las ontologías de dominio especifico modelan las particularidades de las realidades de acuerdo a los propósitos de explotación impuestos [4]

Una evaluación consiste en validar un modelo lo cual implica comprobar que representa de manera fiel el dominio del mundo real. En la actualidad mucha de la población trabaja por medio de la internet, pues es un método que ahorra y facilita varias actividades y la Evaluación de ontologías no es una excepción.

El buen diseño de aplicaciones web se realiza con una correcta elección de una metodología, esta elección depende del contenido que mantendrá al sistema web. El objetivo de este será el facilitar a usuarios de ontologías una evaluación con distintas opciones de muestra de resultados, logrando también que sea amigable a la vista, por tal motivo la elaboración de la aplicación mencionada en este artículo se llevó acabo con la metodología OOHDM, pues es un método que permite especificar el uso de varios meta-modelos especializados: conceptual, navegación y de interfaz de usuario.

2 Metodología OOHDM

La metodología de Método de Diseño Hipermedia Orientado a Objetos, conocida también por sus siglas OOHDM es una metodología que construye aplicaciones web en cuatro pasos o fases. Es un método basado en modelos para el desarrollo de sitios Web, sistemas de información, presentaciones multimediales, quiscos de información, etc. Se trata por separado el diseño de componentes, el modelo de navegación y el diseño de la interfaz. [5] La metodología OOHDM, ha sido utilizada para diseñar diferentes tipos de aplicaciones hipermedia como galerías interactivas, presentaciones multimedia y aplicaciones web [6]. Esta metodología presenta los diagramas de clases navegacionales y el modelo de esquemas de contextos navegacionales que permiten identificar una estructura global de la aplicación [7] El éxito de esta metodología es la clara identificación de los tres diferentes niveles de diseño en forma independiente de la implementación:

1. Diseño conceptual: En este paso se elabora un Modelo Conceptual de dominio de la aplicación utilizando principios de modelado orientados a objetos. OOHDM utiliza el meta-modelo de clases de UML, para expresar el diseño conceptual. El modelo conceptual es representado como un modelo de clases para mostrar el aspecto estático del sistema [8]
   En la Fig. 1 se ilustra el modelo conceptual de la aplicación desarrollada, simplificando en detalle para mejorar su legibilidad destacando las entidades, pero evitando detalles. Este modelo se realizó con el software"StartUM"; software para el diseño de distintos tipos de diagramas usados en la creación de desarrollo software.
2. Diseño navegacional: En esta metodología, el diseño navegacional es construido como una vista sobre el diseño conceptual, admitiendo la construcción de modelos diferentes de acuerdo con los diferentes perfiles de usuario.

   En OOHDM hay una serie de clases especiales predefinidas, que se conocen como clases navegacional: Nodos, Enlaces y Estructuras de acceso, que se organizan dentro de un contexto navegacional;

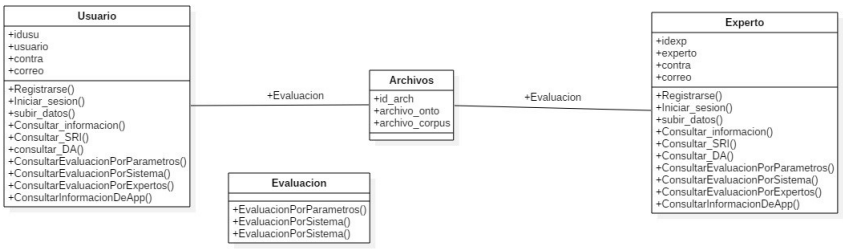

Fig. 1. Diseño conceptual sobre aplicación web para la evaluación de otologías de dominio restringido

- Nodos: Los nodos son contenedores básicos de información de las aplicaciones hipermedia, se define como vistas orientadas a objeto de las clases definidas durante el diseño conceptual usando un lenguaje predefinido, permitiendo así que un nodo sea definido mediante la combinación de atributos de clases diferentes relacionadas en el modelo de diseño conceptual
- Enlaces: Estos reflejan la relación de navegación que puede explorar el usuario, los enlaces son imprescindibles para poder crear vistas diferentes
- Estructuras de Acceso: Las estructuras de acceso actúan como índices o diccionarios que permiten al usuario encontrar de forma rápida y eficiente la información deseada

3. Diseño de Interfaz Abstracta: Es impórtate definir la forma en la cual los objetos navegacional pueden aparecer, de cómo los objetos de interfaz activarán la navegación y el resto de la funcionalidad de la aplicación, que trasformaciones de la interfaz son pertinentes y cuando es necesario realizarlas Modelos de Vistas abstractas de datos (ADVs):

   Los modelos ADVs no son más que representaciones formales que se usan para mostrar: La forma en que se estructura la interfaz; son elementos abstractos en el sentido de que solo representan la interfaz y su dinamismo, y no la implementación, no entra en aspectos concretos como el color de la pantalla o la ubicación en esta de la información. La forma en que la interfaz se relaciona con las clases navegacionales y La forma en que la aplicación reacciona a eventos externos, para ello se usan los ADVs-Charts que son similares a las máquinas de estados, pues a través de esas se pueden indicar los eventos que afectan a una ADV y como esta reacción a ese elemento
4. Implementación: Después de todo lo realizado solo queda llevar los objetos a un lenguaje concreto de programación, para obtener así la implementación ejecutable de la aplicación.

   Cuando se llega a esta fase como diseñador y programador se definen los ítems de información que son parte del dominio del problema, se identifica también, como son organizados los ítems de acuerdo con el perfil del usuario

y su tarea; decidir que interfaz debería ver y como debería comportarse con el fin de implementar todo el entorno de la aplicación Web

3 Aplicación Web para la Evaluación de Ontologías de Dominio Restringido

La aplicación web para Evaluación de Ontologías de Dominio Restringido se implementó siguiendo los diseños antes mencionados, se creó desde un entorno de desarrollo web que permite a programadores servir paginas HTML a Internet, en el caso de este diseño se utilizó HTLM y PHP como lenguajes para implementarla. PHP es un lenguaje de secuencia de comandos de servidor diseñado específicamente para la Web, mientras que MySQL constituye el mejor sistema para la administración de bases de datos relacionales de modo rápido y sólido [9]. El entorno de desarrollo utilizado fue WAMPSERVER; Este en un entorno que proporciona lenguajes de programación adecuados para un desarrollo eficiente, trabaja con MySQL para el caso de Bases de Datos, por lo que la base de datos utilizada se hizo en ese lenguaje (ver Fig. 2), la base de datos tiene como función el control de la cantidad de usuarios que se registran, así como la separación entre usuarios y expertos.

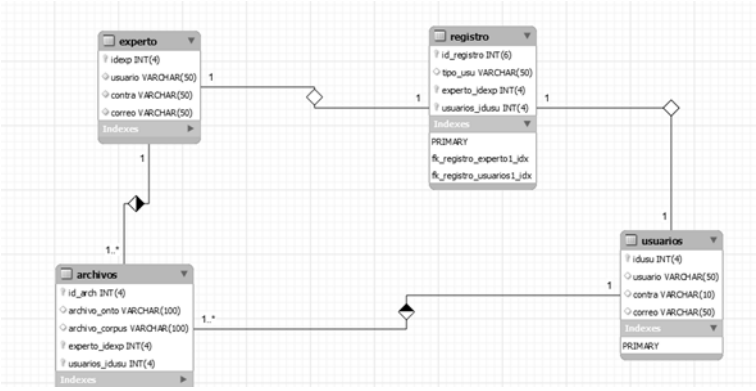

Fig. 2. Diagrama de Base de Datos para Aplicación Web de Evaluación de Ontologías de dominio restringido

Como se mencionó antes el lenguaje utilizado para la dicha aplicación fue HTML y PHP y para lograr mejores interfaces y vistas al usuario se utilizaron hojas de estilo CSS. Para programar los lenguajes se recurrió a DREAMWEAVER; Programa que está destinado a la construcción, diseño y edición de sitios, videos y aplicación web. Cabe mencionar que las imágenes utilizadas se diseñaron por completo en PHOTOSHOP; editor de imágenes.

4 Resultados

Con la ayuda de la metodología y software mencionado se obtuvo una aplicación web con diseño cómodo para el usuario:

– Página Principal: Se muestra el logotipo principal un menú con tres opciones (Página principal, página de servicios que se ofrecen y página de desarrolladores de aplicación), un panel de cambio de imágenes, dos imágenes que llevan al registro e inicio de sesión, ver Fig. 3 y Fig. 4.

Fig. 3. Página principal de Aplicación web.

– Inicio de sesión: Se muestra una ventana de emergencia con opción de dos posibles selecciones (Usuario o Experto) Al seleccionar "Inicio de sesión" se mueve a una ventana con un formulario que solicita el nombre de usuario y la contraseña ver Fig. 5 y Fig. 6.

– Registro: Se muestra un formulario para el llenado de datos ver Fig. 7 (Nombre y Apellido, Correo Electrónico, Usuario, Contraseña, Tipo de usuario).

– Proceso: Se muestran las tres etapas para la evaluación de ontologías ver Fig. 8(Preprocesamiento, Descubrimiento Automáico y Evaluación).

– Preprocesamiento: Se encuentran las tres etapas del Preprocesamiento una llevándote al sitio correcto ver Fig. 9.

– Evaluación: Contiene los tres tipos de evaluación que el sistema ofrece ver Fig. 10.

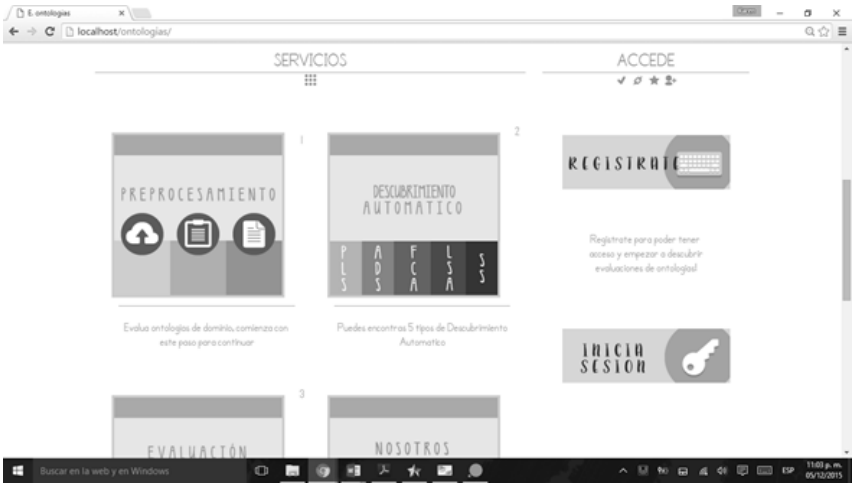

Fig. 4. Segunda vista de página principal de Aplicación web.

Fig. 5. Inicio de sesión para Aplicación web.

Fig. 6. Segunda vista de Inicio de sesión para Aplicación web.

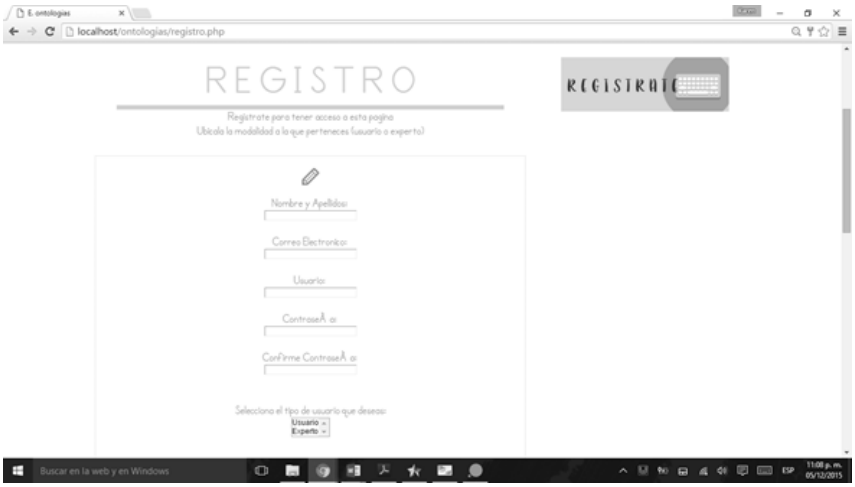

Fig. 7. Registro para entrar en la Aplicación web.

Fig. 8. Etapas para proceso de evaluación de ontologias de dominio para Aplicación web.

Fig. 9. Etapa de Preprocesamiento para el proceso de evaluación de ontologias en Aplicación web.

Fig. 10. Proceso de Evaluación de ontologias en Aplicación web.

5 Conclusiones

El sistema implementado logro ser una aplicación de uso fácil para los usuarios registrados, así como también para expertos del dominio. Es importantes mencionar que la aplicación para Evaluación de Ontologías de Dominio permite también una interpretación gráfica de los resultados obtenidos del sistema de evaluación de ontologías de dominios que se encuentra incrustado en el servidor y que solo permite la visualización en modo texto de los resultados de la evaluación

6 Agradecimientos

Agradezco principalmente el apoyo otorgado por parte de Proyectos de investigación 2015 de la Vicerrectoría de Investigación y Estudios de Posgrado al haber brindado una beca para realizar dicha investigación. Agradezco el apoyo, la oportunidad y dedicación de la Doctora Mireya Tovar Vidal por brindarme la confianza, el tiempo, su animo y los conocimientos adecuados para lograr un buen resultado.

References

1. M. C. J. A, "Tutoria de ontologias," 2008.
2. G. E. Barchini, M. Álvarez, and S. Herrera, "Sistemas de información: nuevos escenarios basados en ontologías," *JISTEM-Journal of Information Systems and Technology Management*, vol. 3, no. 1, pp. 3–18, 2006.

3. M. Tovar, D. Pinto, A. Montes, G. González-Serna, and D. Vilariño, "Evaluación de relaciones ontológicas en corpora de dominio restringido," *Computación y Sistemas*, vol. 19, no. 1, pp. 135–149, 2015.
4. A. C. Ramos, "Metodologas y tecnologas actuales para la construccin de sistemas multimedia," 2015.
5. L. A. Guerrero, "Modelando interfaces para aplicaciones web," *Ingeniería del Software en la Década del*, pp. 227–236, 2000.
6. D. Silva and B. Mercerat, "Construyendo aplicaciones web con una metodología de diseño orientada a objetos," *Revista Colombiana de Computación–RCC*, vol. 2, no. 2, 2001.
7. M. González, S. Abrahão, J. Fons, and O. Pastor, "Evaluando la calidad de métodos para el diseño de aplicaciones web," *I Simpósio Brasileiro de Qualidade de Software*, 2002.
8. D. Schwabe and G. Rossi, "Developing hypermedia applications using oohdm," in *Workshop on Hypermedia Development Process, Methods and Models, Hypertext*, vol. 98, 1998.
9. L. Welling and L. Thomson, *Desarrollo web con PHP y MySQL*. 2005.

Capítulo 10

Grados de conversación y recursos de interacción comunicativa en Twitter

Noemí Elisa Guerrero Contreras

Universidad Nacional Autónoma de México

elisaguerrcon@comunidad.unam.mx

Abstract. Mediante esta investigación se pretende hacer una revisión de la comunicación online a partir del análisis de las características de la interacción comunicativa en las redes sociales en comparación con la conversación cara a cara, tomando como caso específico Twitter que, a semejanza de la modalidad presencial, también ha sido considerada conversacional. En primera instancia, a través de los datos de una serie estudios de caso, esta investigación se enfocará en examinar distintas estrategias como el retuiteo, las menciones (@), la utilización del hashtag, y las respuestas directas (reply) como recursos a través de los cuales los usuarios se integran a una conversación. En segundo término, se buscará contrastar de forma general las propiedades de la comunicación en el marco de la interacción virtual con las de la conversación cara a cara. Por lo tanto, en este trabajo se analizará el cómo los recursos para la comunicación en Twitter se constituyen como prácticas conversacionales.

Keywords. Redes Sociales, Twitter, Conversación en Línea, Recursos de la Comunicación, Comunicación Cara a Cara, Participantes, *Microblogging*, Toma de Turnos.

1 Introducción

Recientemente con la aparición de redes sociales como Facebook, Twitter y otras como Jaiku, Pownce y Yammer ha surgido el interés entre especialistas de diversas disciplinas en la dinámica relativa al funcionamiento y uso de la lengua dentro del marco de estas plataformas en específico.

Sin embargo, algunos otros trabajos, mucho antes de la creación de estos sitios, se han enfocado en el estudio especializado del *computer-mediated discourse* (CMD) al tomar como punto de partida el que exista interacción comunicativa directa entre usuarios a través de mensajes de texto, todo esto dentro del marco comunicativo mediado por el uso de las nuevas tecnologías (computadoras y otros dispositivos inteligentes) [2, 3, 4, 10, 19, 20].

Debido a ello, se han incluido como escenarios salas de chat (Internet Relay Chat), foros virtuales de opinión y otros sitios mediante los cuales los usuarios[1] interactúan en tiempo real.

Asimismo, se han considerado otros medios comunicativos como los blogs[2] para realizar análisis lingüísticos de tipo variacionista [17] y de corte sociolingüístico en Twitter [27][3], Uno de los principales objetivos esta clase de estudios es medir el cambio lingüístico mediante el monitoreo del uso de la lengua por parte de los usuarios de tales sitios a través de un soporte escrito (enriquecido con otros recursos). De manera que el corpus base de estos estudios está constituido por blogs de contenido diverso.

Tanto Facebook como Twitter son plataformas creadas para fomentar la comunicación y la conexión entre amigos y gente conocida con el fin de socializar (*social network sites*), sin embargo, la diferencia más sobresaliente entre estas dos redes parece radicar en que Twitter, como sitio de *microblogging*, se enfoca en ofrecer acceso a información actualizada sobre temas, eventos o personajes de interés para el usuario en tiempo real. Probablemente debido a esto es que dicha plataforma ha sido utilizada con diversos fines además de para los que fue creada.

1.1 Antecedentes

Como ya se ha mencionado, hace ya un tiempo relativamente reciente se han desarrollado diversos trabajos en torno a las redes sociales, la interacción y el impacto de estas principalmente en el ámbito social, sin embargo, algunas también han buscado aproximarse al área del lenguaje y la comunicación.

Trabajos como el de Pano Alamán y Mancera Rueda (2014) abordan el tema de la conversación en Twitter desde una perspectiva del análisis del discurso. Análisis de esta clase se enfocan a estudiar la función de distintas unidades discursivas (marcadores), comúnmente utilizadas en la conversación cara cara, insertas en el contexto virtual que ofrece Twitter. Por otra parte, algunas investigaciones como la de Rossi y Magnani (2012); Huang, Thornton y Efthimiadis (2010); boyd, Golden y Lotar (2010); De Moor (2010) se han concentrado más en el proceso sociotécnico de la dinámica de interacción en Twitter, por lo

[1] La interacción entre usuarios en salas de chat, foros, Facebook, Twitter, etcétera, no necesariamente se limita a dos usuarios, sino que el entorno facilita la intervención múltiple alrededor de un mismo tema de conversación o discusión.

[2] Los *blogs* funcionan como una bitácora virtual que el usuario actualiza de forma regular a través de *posts.* El término *posts* hace referencia a entradas o porciones de contenido que comúnmente se encuentran ordenadas de forma cronológica inversa en la página principal de un *blog*.

[3] La investigación de Reddy, Stanford y Zhong (2014) de tipo sociolingüístico y dialectológico versa sobre procesos de cambio en la lengua y el discurso en Twitter. Un ejemplo de esto es su propuesta de análisis para lo que ellos denominan nuevas formas de acortamientos entre hablantes jóvenes incluyendo términos como *awks* (*awkward*), *adorb* (*adorable*), *ridic* (*ridiculous*), *hilar* (*hilarious*).

que estas propuestas analizan los recursos de interacción que provee esta plataforma a manera de funcionalidades propias de su sistema de organización. La presente investigación pretende profundizar el análisis que merece la dinámica comunicativa en Twitter frente a la interacción conversacional cara a cara sin dejar de lado las características y elementos propios de la plataforma virtual que sirve de marco comunicativo.

A continuación mencionaré los objetivos de esta investigación. Mediante este trabajo se pretende:

- Analizar la comunicación en las redes sociales, tomando como caso específico Twitter, la cual ha sido considerada conversacional. Para acometer esta empresa, en primera instancia el primer punto de atención se enfocará en la dinámica que propicia el uso de Twitter como medio de comunicación para entablar conversaciones virtuales entre varios usuarios y los recursos mediante los que es posible esta clase de interacción.
- En segundo término, se busca contrastar de forma general las propiedades de la comunicación en el marco de la interacción virtual con las de la conversación cara a cara.

A continuación se mostrarán a manera de subtítulos algunos de los principales aspectos de esta investigación.

2 La conversación en Twitter

El contexto conversacional cara a cara y su funcionamiento es muy complejo, puesto que no se limita a un intercambio lingüístico sino que envuelve elementos como la gestualidad, la entonación, el escenario de interacción en el que se desarrolla, el conocimiento cultural de los participantes dentro de una comunidad de habla, entre otros. Por lo que podría parecer desafiante el intentar vincular este marco eventivo con la interacción comunicativa virtual, no obstante, si se examina de cerca, es posible demostrar que las características y recursos propios de la interacción cara a cara no sólo se encuentran presentes sino que se adaptan a un entorno virtual específico.

Twitter como una plataforma cuyo principal objetivo es vincular a la gente, es una red social híbrida puesto que las interacciones *online* que se producen a través este medio no se restringen tan sólo a los tuits sino que también se le permite al usuario compartir su ubicación y contenido multimedia como imágenes y video, de forma directa o a través aplicaciones o enlaces al contenido en tiempo real por lo que, de alguna forma, esto constituye un primer elemento que aproxima a este medio comunicativo a la conversación cara a cara, en el sentido de que mediante estos recursos se captura el contenido del evento desde la *perspectiva* del usuario, además de que se hace uso del soporte escrito.

Otra característica de este sitio de interacción social con respecto al uso de la lengua tiene que ver con que la mayoría de los usuarios promedio[4] utiliza esta plataforma para

[4] Podría establecerse una clasificación de entre los usuarios de *Twitter* de acuerdo al tipo de uso que le dan a su cuenta y al relacionarse con el resto de los usuarios (Honeycutt y Herring, 2009).

conversar con amigos y gente de su entorno social inmediato sobre actividades diarias, asuntos de interés mutuo o simplemente para compartir lo que tienen en mente en ese momento, no obstante a través de la observación y el seguimiento realizado dentro de esta red social es posible afirmar que los usuarios no sólo entablan comunicación con individuos que pertenecen a su círculo social sino que además es muy frecuente que se establezcan conversaciones con usuarios con los que no habían mantenido contacto alguno.

Los datos de la serie de estudios de caso arrojados a través del monitoreo realizado en Twitter han permitido demostrar que las interacciones más extensas y que giran alrededor de temas significativos son las que se dan entre individuos que no habían tenido antes contacto alguno. Al parecer, una de las razones por las que esto sucede tiene que ver con la temática en torno a la que gira una discusión o conversación. Los temas de común interés atraen a los usuarios hacia lo que podríamos denominar una red comunicativa virtual lo que propicia una interacción nutrida que puede convertirse en debate, lo cual también suele suceder en los *weblogs* en los que es posible publicar comentarios de opinión.

Dentro del marco eventivo de la comunicación cara a cara y siguiendo a Paul Grice (1975) existen determinadas condiciones que gobiernan el intercambio comunicativo y la interpretación de este. Las conocidas máximas generales que Grice propuso y que tienen que ver con cantidad, calidad, relevancia y claridad constituyen normas regulativas de la conversación. Por lo tanto, es lógico concluir con que en el intercambio comunicativo virtual es necesario respetar, hasta cierto grado, dichas máximas para lograr éxito comunicativo, sin embargo es evidente que existen diferencias en la proyección de dichos principios en un entorno virtual como el que ofrece Twitter.

Aunado a lo anterior, es importante señalar que si bien la interacción en esta red social tiene un soporte escrito, la mayoría de los usuarios en general produce textos apegados a características propias del registro popular, por lo que este tipo de escritura en línea se caracteriza por ser espontánea, en el sentido de que se produce en tiempo real y no se encuentra regulada dentro de algún marco formal de lengua escrita. Se trata de comunicación orientada a la interacción entre dos o más personas, en consecuencia genera expectativas de intercambio continuo.

Algunos trabajos recientes han buscado enfatizar el carácter dialógico de la lengua escrita en las redes sociales, por lo que se ha utilizado el término *digital networked writing* [6] para referirse a este tipo de escritura. Por otra parte, como ya se mencionó antes, algunos otros investigadores abordan el estudio del uso de la lengua en las redes sociales y otros medios de comunicación *online* desde una perspectiva que pone de relieve el cambio a nivel sociolingüístico [17], aunque es verdad que existen recursos exclusivos para la comunicación a través de internet, también es cierto que se cuenta con evidencia obtenida a través del monitoreo y la investigación[5] que demuestra que de la misma forma es posible plantearlo como expansión de la lengua oral hacia otros medios, como lo es el caso del texto escrito en el contexto de las redes sociales.

[5] Mediante datos que se presentan en otro trabajo, como lo muestra el análisis del uso del verbo *importar* en Twitter.

3 ¿Cuál es la función de los *hashtags*, *retuits* y el símbolo @ como recursos orientados a la interacción comunicativa en Twitter?

En la sección anterior se mencionó que los temas de interés común atraen a más de dos usuarios de Twitter a lo que puede tener el carácter de charla o discusión. A través de esta plataforma tanto el "seguir" las cuentas de otros usuarios como los *hashtags* (#)[6] facilitan la tarea de rastrear un tema de interés (*trending topics*), estos últimos pueden describirse como etiquetas que utilizan el símbolo de # y que sirven para agrupar temas. Su uso comenzó en Twitter pero se ha extendido a otras plataformas como Instagram, Facebook y Vine. A continuación se muestran algunos ejemplos:

#ayotzinapa11meses	#Amici14	#DiaMundialdelMedioAmbiente
#Penabots	#LaVerdadNoEsGuerraSucia	#MetroOceanía
#NationalDonutDay	#ApagónEcológico	#ElCulpableEsElTren

La lista mostrada arriba refleja las tendencias en México, durante un momento y fecha específicos, con relación a distintos temas en Twitter, es decir, la frecuencia de uso de estas etiquetas relativas a determinado tópico en los tuits proyecta la popularidad de una temática. No obstante, debido a que los *hashtags* tienen la función de identificar temas de interés entre los usuarios de esta plataforma también sirven como un recurso para aglomerar a los usuarios entorno a un mismo tema, lo que se asemeja al carácter espontáneo de la conversación cara a cara, la cual no suele verse limitada únicamente a dos participantes (hablante/oyente).

Con relación a este último punto, el trabajo pionero de Erving Goffman (1975, 1976) y el de muchos otros investigadores [8. 9, 25, 26] que abordan conceptos como el de *face work* y *footing* o *participation structure*, siguiendo a Levinson, desde una perspectiva lingüística han subrayado el hecho de que los roles dentro del marco comunicativo no pueden simplificarse tanto como para referir tan sólo a categorías de participantes en las que intervienen hablante, oyente y terceros (presentes o ausentes).

El contexto de la comunicación cara a cara constituye una prueba clara de que es necesario hacer uso de una categorización de roles más fina que refleje los cambios de cada participante de acuerdo a distintas situaciones comunicativas. Una aportación relevante del

[6] Los *hashtags* comenzaron a utilizarse en una plataforma conocida como *Internet Relay Chat* (ICR), la cual comenzó a popularizarse en 1988 y cuya dinámica consistía en el intercambio online de mensajes dirigidos.

trabajo de Goffman, además de que propone desmenuzar los roles de los participantes básicos para desprender de ahí otros más específicos funcionalmente,[7] es que los roles de participación, tanto generales como específicos, no son de carácter fijo o inamovible sino que justamente los papeles más específicos surgen debido a la flexibilidad y situación particular de cada contexto lo que permite cambios en las categorías básicas, por lo tanto dichos roles se complementan.

4 Los Datos

Nuestro corpus consiste en un conjunto de 400 mensajes tomados de la *timeline* pública de distintos usuarios elegidos al azar. Estos tuits fueron recolectados mediante la utilización de Twitter API (v1.1 REST API) (*application programming interface*) durante un lapso de un mes. A partir de los tuits de 174 usuarios se encontró lo siguiente:

- 62.5 % de los mensajes constituyen una respuesta directa (*reply*) a por lo menos uno de los tuits de otro usuario.
- 13.5% de los mensajes escritos pos los usuarios fueron retuiteados de entre 1 hasta 183 veces.
- El 6.7% de los tuits contienen el símbolo de *hashtag* (#).
- 4.5% de los tuits que publicaron los usuarios son mensajes cuya autoría se adjudica a otro usuario, es decir, se trata de retuits.
- El 2.7% de los tuits son preguntas directas a otro usuario.
- 23.2% de los mensajes contienen al menos una mención.

Mediante cada uno de los recursos registrados se configura el entorno conversacional virtual de Twitter. A continuación se presenta una muestra del análisis de los datos realizado.

5 El aspecto conversacional de los recursos utilizados en Twitter

El siguiente ejemplo es un fragmento de una conversación mantenida entre varios usuarios de Twitter y busca ilustrar la propuesta que se sugiere en torno a los recursos para la comunicación en dicha red social como prácticas conversacionales.

[7] Goffman (1981) propone una serie de roles de participación más detallada y específica partiendo de categorías generales tales como *production format* o roles de producción, *participation framework* o roles de recepción ratificados y no ratificados (*over-hearers* o *bystanders* y *eavesdroppers*).

A:@TR/R1[8]_____: Mm qué haría @JoelOrtegaCueva si ésta pierna[9] fuera de su padre, hno o hijo? Es pregunta #MetroOceanía @TapiaFernanda

Serie de respuestas suscitadas:

B: @TR/R2_____: @TR/R1_____ @JoelOrtegaCueva @TapiaFernanda en serio es de un lesionado del metro? Parece corte de arma blanca, creo.

0 retweets 1 favorite

@TR/R3_____: @TR/R1_____ ¿La imagen es de alguna fuente confiable, TR/ R1?

0 retweets 1 favorite

El trasmisor/ receptor 1 responde al TR/R3

@TR/R1_____: @TR/R3_____ sí muy confiable...

0 retweets 1 favorite

@TR/R3_____: @TR/R1_____ ¿Y se puede saber cuál es? Es decir, se agradece la labor periodística, pero no hay que apelar al amarillismo ni al...

0 retweets 0 favorites

TR/R3 Continúa

@TR/R3_____: @TR/R1_____ sentimentalismo, porque si no el trabajo pierde su valor y confiabilidad.

0 retweets 0 favorites

Nueva intervención dirigida a TR/R2

@TR/R4_____: @TR/R2_____ @TR/R1_____ @JoelOrtegaCueva @TapiaFernanda / Arma blanca? Como de qué tamaño?

1 retweet 0 favorites

TR/R2 Contesta a TR/R4

@TR/R2_____: @TR/R4_____ @TR/R1_____ @JoelOrtegaCueva @TapiaFernanda mi señor con que tenga filo no necesita ser grande. Pero es mi opinión o pregunta.

1 retweet 0 favorites

Nueva intervención dirigida a TR/R1

@TR/R5_____: @TR/R1_____ @JoelOrtegaCueva @TapiaFernanda Acusaría a @m_ebrard del incidente. Falta mantenimiento, igual que en la L12 y no lo entiende

8 retweets 3 favorites

---

[8] Las iniciales TR/R se utilizan para señalar a cada participante sencillamente como transmisor/receptor. Sólo se omitió el nombre de los usuarios que interactúan en la conversación. Los nombres de usuarios que sí se hacen explícitos refieren sólo a "menciones" de estos.

[9] El contenido de este *tuit* incluye una imagen del evento que describe el mensaje.

TR/R1 retoma la conversación que mantenía con TR/R3
@TR/R1_____: @TR/R3_____ no TR/R3, la fuente no me autorizó difundir quién es...Y no es amarillismo, es lo que sucedió a una víctima. Es realidad.
1 retweet 1 favorite
Nueva intervención dirigida a TR/R1
@TR/R6_____: @TR/R1_____ Que fuerte!
0 retweets 0 favorites
TR/R3 responde a TR/R1
@TR/R3_____: @TR/R1_____ Bien, sólo es una suge-rencia. Sigo tu trabajo en el programa de Fernanda y me parece muy loable, pero en esa profesión...
0 retweets 0 favorites
TR/R3 continúa
@TR/R3_____: @TR/R1_____ hay que ser justamente eso, profesionales.
0 retweets 0 favorites
TR/R1 entabla conversación a partir de los comentarios hechos por TR/R5
@TR/R1_____: @TR/R5_____ @joelortegacueva @tapia-fernanda @m_ebrard seguro después de esto le darán otro "respetuoso y honorable" cargo #ElCulpableEsElTren
5 retweets 1 favorite
Nueva intervención dirigida a TR/R1
@TR/R6_____: @TR/R1_____ @JoelOrtegaCueva @Tapia-Fernanda las lesiones de este tipo de accidentes son horrendas y más si la colisión fue tan brutal.
0 retweets 0 favorites
Nueva intervención dirigida a TR/R1 y TR/R3 retomando el asunto discutido entre estos dos últimos participantes.
@TR/R7_____: @TR/R1_____ @TR/R3_____ amarillismo? Yo no lo veo en ningún lado. En otros medios sol hablan de lesionados este señor de la foto....
0 retweets 0 favorites
@TR/R7_____: @TR/R1_____ @TR/R3_____ ya quedo marcado de por vida salvo que tenga una buena terapia que cuesta dudo que quede bien
0 retweets 0 favorites
@TR/R7_____: @TR/R1_____ @TR/R3_____ y la pregunta es @JoelOrtegaCueva @joelortse hará cargo de la terapia así como indemnización?....

0 retweets 0 favorites
@TR/R7_____: @TR/R1_____ @TR/R3_____ o también @JoelOrtegaCueva culpara a @m_ebrard @marcelde esto
0 retweets 0 favorites
TR/R3 responde a TR/R7
@TR/R3_____: @TR/R7_____ No lo sé, creo que es una interrogante a la cual no puedo responder yo, y creo que ya no es pertinente que me menciones.
0 retweets 0 favorites
Nueva intervención dirigida a TR/R1
@TR/R8_____: @TR/R1_____ @jorgearomog @JoelOrtegaCueva @TapiaFernanda buscaría otro hueso porque vivir fuera del presupuesto es vivir en el error
1 retweet 1 favorite
Nueva intervención dirigida a TR/R1
@TR/R9_____: @TR/R1_____ @_martinmoreno @JoelOrtegaCueva @TapiaFernanda Se le salió el mondongo ...
0 retweets 0 favorites
Nueva intervención dirigida a TR/R1
TR/R10_____:@TR/R1_____ @TapiaFernanda @_martinmoreno Lo bueno es que no hay heridos de gravedad.
0 retweets 0 favorites
Nueva intervención dirigida a TR/R1
TR/R11_____: @TR/R1_____ @_martinmoreno @JoelOrtegaCueva @YuririaSierra como los jóvenes mutilados de venezuela #Ocenía pic.twitter.com/cQeQRZYA0g[10]
0 retweets 0 favorites
Nueva intervención
TR/R12_____:@TR/R1_____ @JoelOrtegaCueva @TapiaFernanda No que no había lesionados! Entonces a qué le llaman lesión?
0 retweets 0 favorites
Nueva intervención
TR/R13_____: @TR/R1_____ @_martinmoreno @JoelOrtegaCueva @TapiaFernanda donde anda aquel? Que responda por el accidente!!!!!
0 retweets 0 favorites
Nueva intervención dirigida a TR/R5 y TR/R1

[10] Enlace de imagen

@TR/R14_____:@TR/R5_____ @TR/R1_____ @JoelOr-
tegaCueva @TapiaFernanda también le cargaran ese muerto a
@m_ebrard? Hace cuanto que dejó el cargo?
0 retweets 0 favorites
Nueva intervención dirigida a TR/R1
TR/R14_____: "@TR/R1_____: qué haría @JoelOrtega-
Cueva si ésta pierna fuera de su padre o hijo? pic.twit-
ter.com/EUrffebXG7"[11] créelo que no usan el metro.
0 retweets 0 favorites
La conversación continúa…

El fragmento arriba citado es sólo un ejemplo del nivel conversacional que descubrimos en Twitter. Como acaba de advertirse, la interacción no se limita a mensajes de texto, puesto que la primera intervención se vale de una fotografía que muestra de manera explícita la situación referida por el usuario. Si se sigue el hilo de esta comunicación es sencillo darse cuenta de que esta intervención funge como 'disparador' de la conversación, ya que se trata de una pregunta directa que el usuario 'lanza' a sus "seguidores" en la red, quienes a su vez responden (Fig.1).

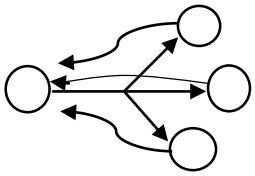

Fig 1. Esquema comunicativo online.

Dicha cuestión suscita una respuesta casi inmediata, ya que el tuit que abre la conversación fue publicado el 4 de mayo de 2015 a las 9:43 PM y a las 9:47 PM de la misma fecha se dio la primera respuesta. La conversación a la que el usuario dio inicio y en la que se mantuvo activo interactuando con otros usuarios en la red se dio de las 9:43 PM a las 11:40 PM, sin embargo, debido a que los mensajes permanecen en la *timeline*[12] del usuario pueden ser vistos en cualquier momento después de que fueron emitidos.

[11] La intervención de este nuevo participante en la conversación para dar su opinión al respecto incluye lo que podría considerarse una cita directa del primer *tuit*.

[12] El término *timeline* refiere la página principal de Twitter de cada usuario en la que se encuentran todos los mensajes que ha emitido ordenados de forma cronológica.

5.1 El retuiteo.

Otro recurso mediante el cual otros usuarios pueden verse involucrados en la conversación de forma indirecta es mediante el retuit de un mensaje, esto implica el que otro usuario, que sí participa en la conversación, reenvíe los comentarios de uno o más participantes a sus seguidores, quienes tienen la misma función que los *bydroppers* en la conversación cara cara, esto permite que otros usuarios que no siguen al que dio inicio a la conversación, y que por tanto son ajenos a esta, también puedan participar.

Lo anterior sucede en el caso del fragmento presentado, puesto que nuevas intervenciones se dieron en torno al mismo tema, aún un día después de que se produjo la primera invitación a la interacción, a partir de las 8:56 AM del 5 de mayo. Entonces, entre más se retuitée[13] un mensaje es mayor la posibilidad de incluir a más usuarios como participantes potenciales en la conversación.

5.2 El uso del símbolo @.

Con relación al fragmento arriba presentado es importante señalar que la función del símbolo @ en Twitter, el cual refiere al usuario y se utiliza para hacer mención o referencia explícita y directa a un nombre de usuario en específico, lo cual demuestra que en los mensajes de las conversaciones mantenidas en esta red pueden encontrarse dirigidos o diseñados de manera particular para otros usuarios. Entonces, en primer lugar el uso del símbolo @ constituye un recurso para la interacción dirigida, puesto que en la comunicación cara a cara equivale a dirigir la palabra a un interlocutor determinado.

Además de ello, la conversación de Twitter antes mostrada nos deja ver que el signo @ se utiliza también para hacer referencia indirecta a algún usuario, aunque no esté involucrado en la conversación, el cual tiene el mismo valor que el participante con el rol de *indirect target* mencionado por Levinson (1987). Un ejemplo son las referencias de los múltiples participantes de la conversación a usuarios como @JoelOrtegaCueva, @TapiaFernanda, @m_ebrard, @_martinmoreno y @YuririaSierra.

Los usuarios anteriormente mencionados no participan en ningún momento en la conversación, sin embargo son referidos por los participantes activos. Esto indica que este tipo de mención de usuarios no involucrados en el intercambio comunicativo puede equipararse en la interacción cara a cara al acto de hablar de alguien que no está presente con nuestro interlocutor, debido a que es pertinente o relevante de alguna manera para lo que se busca comunicar.

En el caso del fragmento arriba mostrado, es posible entender porque una conversación que gira en torno a temas como #MetroOceanía y #ElCulpableEsElTren se vincula con

[13] El fragmento de conversación en Twitter arriba presentado muestra un total de 16 retuits distribuidos entre 5 mensajes distintos, lo que amplía el alcance de la conversación a un mayor número de usuarios.

usuarios de Twitter como @JoelOrtegaCueva, es decir, es factible establecer los vínculos comunicativos necesarios.

El desarrollo de la conversación permite apreciar que el usuario que la inicia, a semejanza de lo que sucede en la interacción cara a cara, elige responder a quien desea (dirige su mensaje), puesto que ignora el primer comentario (tuit) y el cuestionamiento que se le plantea en respuesta, tal vez porque parece desacreditar de alguna manera su participación, y responde a un tercer usuario (TR/R3) que interviene en la comunicación.

De igual manera, el TR/R1 propicia el flujo de la conversación al responder al comentario de un nuevo usuario. Esto demuestra que los usuarios con el rol de *bydroppers* pueden cambiar su posición y convertirse en participantes activos en la conversación.

Nueva intervención dirigida a TR/R1

@TR/R5_____: @TR/R1_____ @JoelOrtegaCueva @TapiaFernanda Acusaría a @m_ebrard del incidente. Falta mantenimiento, igual que en la L12 y no lo entiende

TR/R1 entabla conversación a partir de los comentarios hechos por TR/R5

@TR/R1_____: @TR/R5_____ @joelortegacueva @tapiafernanda @m_ebrard seguro después de esto le darán otro "respetuoso y honorable" cargo #ElCulpableEsElTren

Otra estrategia utilizada que se observa en la interacción en Twitter arriba expuesta es la de citar directamente el primer mensaje con el fin de retomar el hilo de la conversación y responder a la pregunta inicial, lo cual hace el TR/R14.

Nueva intervención dirigida a TR/R1

TR/R14_____: "@TR/R1_____: qué haría @JoelOrtegaCueva si ésta pierna fuera de su padre o hijo? pic.twitter.com/EUrffebXG7"[14] créelo que no usan el metro.

En la conversación de Twitter antes mostrada es posible notar que el usuario autor (TR/R1) propicia la conversación mediante una pregunta directa y participa activamente a lo largo de la discusión. Más adelante, se puede advertir que el TR/R1 contesta a diferentes usuarios que intervienen dando su opinión sobre el asunto en respuesta a esta primera intervención. Sin embargo, al final de este fragmento puede notarse que también se producen intervenciones que no se encuentran dirigidas a ningún participante.

Este tipo de tuits equivalen a lo que en la conversación cara a cara se interpreta como una exposición de ideas desde la perspectiva del transmisor/receptor, es decir al acto de

[14] La intervención de este nuevo participante en la conversación para dar su opinión al respecto incluye lo que podría denominarse como una cita directa del primer tuit.

externar una opinión personal (la audiencia). Adelante se repiten un par de ejemplos que ilustran lo anteriormente mencionado:

Nueva intervención
TR/R12_____:@TR/R1_____ @JoelOrtegaCueva @Tapia-Fernanda No que no había lesionados! Entonces a qué le llaman lesión?
Nueva intervención
TR/R13_____: @TR/R1_____ @_martinmoreno @JoelOrtegaCueva @TapiaFernanda dónde anda aquel? Que responda por el accidente!!!!!

Las intervenciones antes mostradas dejan ver que a pesar de que estos tuits se encuentran formulados dentro del esquema de respuesta (*reply*) en esta plataforma, debido a que al inicio del mensaje se hace mención del usuario autor (TR/R1), en realidad ninguno de los dos mensajes se ajustan al perfil de contestación a la pregunta inicial ni a ninguna otra participación, sino que más bien pueden interpretarse como posturas personales relativas al tema.

Los tuits pueden estar dirigidos a más de un usuario como lo muestra el siguiente ejemplo:

Nueva intervención dirigida a TR/R1 y TR/R3 retomando el asunto discutido entre estos dos últimos participantes.
@TR/R7_____: @TR/R1_____ @TR/R3_____ amarillismo? Yo no lo veo en ningún lado. En otros medios solo hablan de lesionados este señor de la foto....
@TR/R7_____: @TR/R1_____ @TR/R3_____ ya quedo marcado de por vida salvo que tenga una buena terapia que cuesta dudo que quede bien.

Esta participación muestra a un nuevo usuario que interviene en la comunicación que habían mantenido el usuario que comienza la conversación y el TR/R3 al inicio y da su opinión al respecto, de manera que sí puede considerarse como una respuesta dirigida. La toma de turnos virtual es mucho más libre que en la conversación cara a cara en el sentido de que cada participante decide si va a responder, cuándo y cómo lo hará, de tal forma que no es necesario que la respuesta sea inmediata como sucede en el marco comunicativo tradicional.

Algunas interacciones en esta red social son distintas a la ya presentada, puesto que pueden describirse de la siguiente forma: el tuit inicial funciona como una especie de "anzuelo" flotante en la red, mediante el cual el usuario busca ocasionar reacciones diversas en otros usuarios, no obstante la participación a lo largo de la discusión del usuario que "lanzó el anzuelo" es casi nula (*broadcasting communication*).

6 Conclusiones

En síntesis, uno de los objetivos de este trabajo consistió en demostrar que dentro de un marco situacional virtual como el que proporcionan plataformas de microblogging como Twitter se producen interacciones comunicativas con un nivel de complejidad propio, por lo que constituyen un tipo de comunicación que merece atención en sí mismo. De igual forma, se mostró que existen diversos recursos especializados para la comunicación virtual que pueden equipararse a los de la conversación cara a cara.

El ejemplo de la conversación antes mostrado expuso con claridad el hecho de que la comunicación que se produce en línea comparte rasgos y recursos con la que se da cara a cara, aunque ambas poseen características distintivas adaptadas al entorno en el que se producen y desarrollan, no obstante no debe pasarse por alto el hecho de que en ambas situaciones se trata de intercambio comunicativo y, en el caso específico de Twitter, dicho intercambio se encuentra mediado por un soporte escrito fundamentado en la oralidad.

La cuestión de la multimodalidad en la conversación cara a cara (gestualidad entre otros recursos comunicativos) puede encontrar cauce en el uso y la función del emoji como lenguaje tanto en esta plataforma como en otras, sin embargo consideramos que es tema que merece un estudio aparte y que por cuestiones de espacio no será tratado aquí.

Referencias

1. Androutsopoulos, J. Language change and digital media: a review of conceptions and evidence. In Nikolas Coupland y Tore Kristiansen, editors, *Standard Languages and Language Standards in a Changing Europe*. Novus, Oslo (2011)
2. Baron, N. S. Computer Mediated Communication as a Force in Language Change. *Visible Language* 18(2): 118-141 (1984)
3. Baym, N. The Emergence of Community in Computer-Mediated Communication. In: Jones, S. (ed.), *Cybersociety: Computer Mediated Communication and Community*. Thousand Oaks (1996)
4. Bays, H. 1998. Framing and face in Internet eschanges: A socio-cognitive approach. *Linguistik Online, 1*(1) (1998).
5. boyd, d.; Golder, S., and Lotan. "Tweet, Tweet, Retweet: Conversational Aspects of Retweeting on Twitter." *HICSS-43*. IEEE: Kauai, HI, January 6 (2010).
6. boyd, d. Social network sites as networked publics. Affordances, dynamics, and impli-cations. In: Z. Papacharissi (ed.) *A networked self. Identity, community, and culture on social network sites*. New York, London: Routledge.39-58 (2011)
7. boyd, d. and Crawford, K. Critical questions for big data. *Information, Communication & Society*, 15(5):662–679, May (2012)
8. Brown, P., and Levinson, S. C. Universals in language usage: Politeness phenomena. In E. N. Goody (Ed.), *Questions and politeness: strategies in social interaction* (pp. 56-311). Cambridge University Press (1978)

9. Brown, P., and Levinson, S. C. *Politeness: Some universals in language usage*. Cambridge University Press (1987)
10. Cherny, L. The Modal Complexity of Speech Events in a Social MUD to appear in *Electronic Journal of Communication*, Summer (1995)
11. Crystal, D. *Internet Linguistics*. London_ Routledge (2011)
12. De Moor, A. Conversations in Context: A Twitter Case for Social Media Systems Design. In *I-SEMANTICS '10 Proceedings of the 6th International Conference on Semantic Systems*. 29 (2010)
13. Cunha, E.; Magno G.; Comarela, G.; Almeida, V.; Gonçalves, M. and Benevenuto, F. Analyzing the dynamic evolution of hashtags on twitter: a language-based approach. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*, pages 58–65, Portland, Oregon, June. ACL (2011)
14. Goffman, E. *Frame Analysis: An Essay on the Organization of Experience*. Cambridge, Mass.: Harvard University Press (1975).
15. Goffman, E. Replies and Responses. *Language in Society* 5:3, pp. 257-313 (1976).
16. Goffman, E. *Forms of talk*. Philadelphia: University of Pennsylvania Press (1981)
17. Grice, H.P. 'Logic and conversation' In Cole, P. & Morgan, J. (eds.) *Syntax and Semantics*, Volume 3. New York: Academic Press. pp. 41-58 (1975)
18. Grieve, J.;Biber, D.; Friginal, E., and Nekrasova, T. Variation among blog text types: A multi-dimensional analysis. In Alexander Mehler, Serge Sharoff and Marina Santini (editors) *Genres on the Web: Corpus Studies and Computational Models*. New York: Springer-Verlag (2010)
19. Haase, M. et al. Internetkommunication und Sprachwandel. *Sprachwandel durch Computer*. Institut für semantische Informationsverarveitung, Universität Osnabrück. http://link.springer.com/chapter/10.1007/978-3-322-91416-3_3#page-1 (1997) Accedido el 26 de mayo de 2015.
20. Herring, S. C. Interactional Coherence in CMC. *Journal of Computer-Mediated Communication.* 4, 4 (1999).
21. Herring, S. C. Language and the Internet. In W. Donsbach (Ed.), *International Encyclopedia of Communication* (pp. 2640-2645). Blackwell Publishers (2008)
22. Honeycutt, C., and Herring, S. Beyond Microblogging: Conversation and Collaboration in Twitter. *Proc 42nd HICSS*, IEEE Press (2009)
23. Hu, Y.; Talamadupula, K., and Kambhampati, S. Dude, srsly? : The surprisingly formal nature of twitter's language. In *Proceedings of ICWSM* (2013)
24. Huang, J., Thornton, K. M. y Efthimiadis, E. N. Conversational Tagging in Twitter. In HT '10 *Proceedings of the 21st ACM conference on Hypertext and hypermedia* (2010)
25. Huberman, B.; Romero, D.; and Wu, F. Social Networks that Matter: Twitter Under the Microscope. *First Monday*. 14(1) (2009).
26. Java, A.; Song, X.; Finn, T.; and Tseng, B. Why we Twitter: Understanding microblogging usage and communities. *Proc. Joint 9th WEBKDD and 1st SNA-KDD Workshop*, ACM Press (2007)
27. Kelly, Ryan. *Twitter Study Reveals Interesting Results About Usage*. San Antonio, Texas: Pear Analytics. (2009)

28. Lakoff, G. Humanistic Linguistics. In Francis P. Dineen (ed) *Georgetown Roundtable on Languages and Linguistics*. Georgetown University Press (1973).
29. Levinson, S. C. Minimization and conversational inference. In M. Bertuccelli Papi, & J. Verschueren (Eds.), The pragmatic perspective: Selected papers from the 1985 International Pragmatics Conference (pp. 61-129). Benjamins (1987)
30. Pano Alamán, A. y Mancera Rueda, A. La "conversación" en Twitter: las unidades discursivas y el uso de marcadores interactivos en los intercambios con parlamentarios españoles en esta red social. En *Estudios de Lingüística del Español* 35.1, pp. 234-268, (2014)
31. Reddy, S.; Stanford, J.; and Zhong, J. A Twitter-Based Study of Newly Formed Clippings in American English. In the *Annual Meeting of the American Dialect Society* (ADS) (2014)
32. Rossi, L. and Magnani, M. Conversation Practices and Network Structure in Twitter. In *Proceedings of the Sixth International AAI Conference on Weblogs and Social Media* (2012)

Índice de Autores

Compiladores

Mireya Tovar Vidal
Beatriz Beltrán Martínez
Karen Leticia Vazquez Flores

Revisores

Hilda Cartillo Zacatenco
Claudia Zepeda Cortes
Meliza Contreras
Mireya Tovar Vidal
José Alejandro Reyes Ortiz
María Josefa Somodevilla García
Darnes Vilariño Ayala
Beatriz Beltrán Martinez
David Eduardo Pinto Avendaño

Editores

Mireya Tovar Vidal
Darnes Vilariño Ayala
Beatriz Beltrán Martínez

Tendencias en la Ingeniería del Lenguaje y del Conocimiento

Se terminó de editar en noviembre de 2016 en la Facultad de Ciencias de la Computación de la Benemérita Universidad Autónoma de Puebla, Av. San Claudio y 14 Sur, Ciudad Universitaria, Puebla, Puebla, C.P. 72592.

El Cuidado de la Edición es de:

Mireya Tovar Vidal
Darnes Vilariño Ayala
Beatriz Beltrán Martínez

Se reproducen 250 CD's
Peso del Archivo PDF: 3.66 MB