Empresa 2.0:
Detección de plagio y análisis de opiniones

Enrique Vallés Balaguer

Departamento de Sistemas Informáticos y Computación

Directores:
Paolo Rosso, Universidad Politécnica de Valencia, España
Viviana Mascardi, Università di Genova, Italia

Tesis desarrollada dentro del Máster en Inteligencia Artificial,
Reconocimiento de Formas e Imagen Digital

Valencia, junio de 2010

Resumen

La llegada de la Web 2.0 ha supuesto una auténtica revolución dentro del mundo empresarial. Las empresas que mejor se han adaptado a este nuevo sistema se han convertido en las empresas de referencia en la actualidad. Y es que la Web 2.0 ha sido toda una revolución, ya que ha proporcionado un mayor contacto entre empresa y clientes mejorando la relación entre éstos. Una de las grandes ventajas que puede aprovechar una empresa de la Web 2.0 son las opiniones que comparten los consumidores sobre marcas y productos, las cuales marcarán las tendencias del mercado.

Sin embargo, la Web 2.0 ha ampliado varios problemas que una empresa debe solventar. La facilidad de acceso a la información que proporciona la Web ha propiciado un aumento en los casos de plagio entre empresas. Las empresas deben protegerse ante aquellas empresas desleales que aprovechan las ideas de otros para atribuirse un mérito ajeno.

En este trabajo describimos el estado del arte de la detección automática del plagio. Además, presentamos unos experimentos realizados con una herramienta para detección de plagio disponible en la Web con la que una empresa puede protegerse ante infracciones en su propiedad intelectual. Demostraremos con estos experimentos la dificultad que entraña para una empresa protegerse del plagio con las herramientas disponibles actualmente lo que acentúa la necesidad de estudiar nuevos métodos automáticos para la protección del plagio para las empresas.

A continuación y siguiendo con la línea de estudiar las posibilidades que tiene una empresa en la Web 2.0, describimos el estado del arte del análisis de opiniones. A continuación proponemos un algoritmo para el análisis de opiniones y la posterior intercambio de información entre empresas. Para comprobar la eficacia de dicho algoritmo presentamos unos experimentos en el dominio del turismo.

Abstract

The advent of Web 2.0 has brought a revolution in the business world. Enterprise that have adapted to this new system, have become the leading enterprises today. Web 2.0 has been a revolution, as it provided greater contact between enterprise and customers by improving the relationship between them. One of the great advantages that a enterprise can take advantage of Web 2.0 are the opinions shared by consumers about brands and products, which will set the market trends.

However, Web 2.0 has expanded several problems that a enterprise must solve. The Web has provided ease of access to information, but this has increased cases of plagiarism among enterprises. Enterprises must protect their ideas of the unfair business, which take advantage of the ideas of others to ascribe merit of others.

In this work, we describe the state of the art of automatic plagiarism detection. We also present some experiments with a plagiarism detection tool available on the Web with which a enterprise can protect its intellectual property. These experiments demonstrate the difficulty for a enterprise to protect its intellectual property with currently available tools which stresses the need to explore new approaches to automatic plagiarism detection.

Then, following the line of studying the possibilities for a Web 2.0 enterprise, we describe the state of the art of opinion mining. Here we propose an algorithm for the analysis of opinions and the subsequent exchange of information between enterprises. To verify the effectiveness of the algorithm we present some experiments in the domain of tourism.

Agredecimientos

En primer lugar quiero dar mi agradecimiento a Paolo Rosso por su dirección en esta tesis, por sus importantes consejos y su paciencia.

Parte del trabajo de investigación de esta tesis, se ha llevado a cabo durante una estancia de 4 meses en la Universidad de Génova (Italia). Es por ello que debo expresar mi más sincero agradecimiento a la profesora Viviana Mascardi y a la doctora Angela Locoro, que durante mi estancia en Génova aportaron valiosos consejos y observaciones a la tesis.

Me gustaría agradecer también, a los miembros del grupo *Natural Language Engineering Lab (NLE Lab)*, al cual pertenezco, y muy especialmente a Alberto Barrón y Antonion Reyes, por toda la ayuda y consejos que me ofrecieron para la realización de esta tesis.

Este trabajo no hubiera sido posible sin el apoyo incondicional de mi familia. Por ello quiero agradecer a mis padres y a mi hermana Àngels, por haber estado siempre a mi lado.

Esta tesis se ha desarrollado en el marco del proyecto del MICINN (Plan I+D+i): *TEXT-ENTERPRISE 2.0: Técnicas de Comprensión de textos aplicadas a las necesidades de la Empresa 2.0* (TIN2009-13391-C04-03)

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Con la llegada de la Web 2.0 las empresas han tenido que amoldarse a las nuevas tecnologías. Las empresas han visto como gracias a la Web 2.0, las relaciones con los clientes han mejorado mucho, ya que el contacto con estos es mucho más directo. Sin embargo, aunque la Web 2.0 ha aportado grandes ventajas, también se han visto algunas desventajas.

En este trabajo hemos tenido la intención de abordar algunos de los problemas con los que se enfrenta una empresa moderna: la protección ante el plagio y la recopilación de información a través de las opiniones de los consumidores realizadas en la Web 2.0.

1.1. Descripción del problema

Las empresas son conocedoras de que el éxito empresarial está estrechamente relacionado en ofrecer al consumidor productos y servicios que sean de su agrado y, por otro lado, introducir novedades dentro de las tendencias del mercado. Es decir que, tanto los productos y servicios ofertados como las novedades introducidas han de cumplir varios factores: deben de cubrir las necesidades de los consumidores, tienen que estar dentro de las tendencias del mercado, y sobretodo deben ser del agrado de los consumidores.

Es por esto que no resulta extraño que entre las características de las empresas con alto crecimiento se encuentren la implementación de nuevas tecnologías y elementos referentes al marketing como la identificación de mercados para el producto o desarrollo de productos para clientes existentes, el desarrollo de nuevos mercados, la ampliación de la base de clientes y la aplicación de una estrategia competitiva de diferenciación de producto con enfoque en el mercado [103]. En [112], Storey reafirma esta idea y destaca como una de las necesidades para el crecimiento de las empresas, entre otras, la innovación de productos.

Actualmente, las empresas que han apostado por el marketing en los medios sociales, como blogs y redes sociales, son las que mayores ventajas han conseguido en un mercado competitivo, y cada vez más exigente. No obstante, no todos los aspectos de los medios

sociales son positivos. Si por una parte, los medios sociales permiten a las empresas tener un mayor contacto con el consumidor informándole de sus productos, sus servicios, sus ideas, así como sus novedades; por otro lado, esta información no sólo esta al alcance de la mano para los consumidores, sino que también es visible para las empresas competidoras. Por desgracia, existen empresas que utilizan dicha información de forma muy poco ortodoxa, puesto que la utilizan para sus propios fines, es decir, copian productos, servicios e incluso ideas de otras empresas. Por este motivo, las empresas están obligadas a protegerse de aquellas empresas que infringen la propiedad intelectual ajena.

Sin embargo, las ventajas que aportan los medio digitales a las empresas son mucho mayores que las desventajas. Sumando a la ventaja anteriormente comentada en la que las empresas tienen un contacto más estrecho con los consumidores, existe otro punto a su favor no menos importante: los consumidores comparten a través de los medios sociales sus opiniones sobre productos y marcas. Conseguir analizar estas opiniones es de vital importancia para el éxito de una empresa. Esto es debido a que las empresas se enfrentan con un duro problema para conseguir que los productos se ajusten a las necesidades y los gustos de los consumidores. Por tanto, las opiniones que comparten los consumidores en las redes sociales, tanto de los productos propios como de los productos de los competidores, puede ser de infinita ayuda para mejorar los productos adaptándolos a las tendencias del mercado.

1.2. Plagio en las empresas

Internet representa uno de los mayores avances en la historia en materia de comunicación. Gracias a ella se puede acceder de forma inmediata a una gran cantidad de información, sin importar las distancias. Sin embargo, la facilidad al acceso a la información, ha incrementado el número de casos en los que se comete plagio. Si bien es cierto que el problema del plagio es tan antiguo como la misma historia (ya en el siglo V a.C. en un certamen de poesía, varios concursantes presentaron como propias, viejas obras existentes en la biblioteca de Alejandría y que, descubiertos, se les sancionó como ladrones [48]), Internet ha abierto definitivamente las puertas de par en par. La gran cantidad de información y la facilidad de acceso a ella, es una tentación demasiado grande para poder resistirse ante una falta de inspiración; según Janett [34]:

> *El escritor novato frecuentemente se bloquea frente a una hoja en blanco y se siente tan inseguro que en ocasiones, por facilismo o por ignorancia, toma sin ningún reparo como propias, ideas, frases o párrafos de un documento y se atribuye la autoría del mismo.*

El plagio se ha convertido, tras la aparición de las redes sociales, en uno de los mayores problemas de la sociedad. Existe la falsa idea de que se puede utilizar con total libertad el material que se publica en la Web por el hecho de ser esta red de dominio público [79]. Sin embargo, la definición de plagio no es ambigua y no deja lugar a dudas:

> *Se entiende como plagio la apropiación, presentación y utilización de material intelectual ajeno, sin el debido reconocimiento de su fuente original. Constituye, por lo tanto, un acto fraudulento, en el cual existe presunción de intencionalidad, en el sentido de hacer parecer un determinado conocimiento, labor o trabajo, como producto propio; y de desconocer la participación de otros en su generación, aplicación o en su perfeccionamiento* [97].

Por tanto, el ánimo de engañar subyace en la acción de plagio con propósitos que van desde la simple intención de un estudiante de obtener una buena calificación en un trabajo, pasando por la exaltación personal o encumbramiento personal de un científico, o incluso, para obtener un desproporcionado lucro o prestigio dentro del mundo empresarial. En definitiva, el plagio es un acto deleznable porque no sólo conlleva el robo de una idea o trabajo, sino que ésta no puede separarse del esfuerzo y mérito del autor original. Y es que la correcta atribución de autoría corresponde al esfuerzo, al mérito y a las capacidades de quien generó la información o el conocimiento y, en tal asignación, que sólo debe alcanzar a las personas que la merecen [1].

Si bien es cierto que a nivel académico es donde mayor incremento de números de plagio se ha producido (según un estudio de Kloda y Nicholson, uno de cada tres estudiantes canadienses afirman haber cometido alguna vez plagio [57]), no sólo se limita a este ámbito. Todo lo contrario, el plagio es frecuente en todas las áreas imaginables:

- **Académico**: "El rector de la Universidad de Chile reconoce que el decano de derecho, Roberto Nahum, plagió la tesis de un alumno" [1]
- **Literario**: "Carmen Gómez Ojea estudia denunciar por plagio al último Premio Nadal" [2]
- **Político**: "El partido de Calderón en México copia un anuncio del PSOE" [3]
- **Científico**: "La revista *Research Policy* se ha visto obligada a retirar de su hemeroteca un trabajo del año 1993 firmado por un profesor alemán, Hans Werner Gottinger. El estudio se parecía sospechosamente a otro del *Journal of Business* datado en 1980" [4]

E incluso el plagio afecta al mundo empresarial. Y es que, aunque la Web ha aportado grandes ventajas a las empresas, como un contacto mayor con el consumidor, también tiene su parte negativa. Internet está produciendo un aumento en conductas de competencia desleal. Así, se dan casos de confusión o imitación, como la creación de dominios similares a otros más conocidos. Por ejemplo, el dominio *microsf.com* buscaba el parecido con *microsoft.com* [79]; o el plagio de páginas Web.

[1] http://www.infinita.cl/ (Radio Infinita, 2009)
[2] http://www.elcomerciodigital.com/ (El comercio digital, 2008)
[3] http://www.elpais.com/ (El País, 2009)
[4] http://www.elmundo.es (El Mundo, 2007)

Es por esta razón que las empresas deben de buscar un modo de proteger su información para que los competidores no utilicen su información de forma incorrecta. Pero, debido a la expansión de la Web 2.0 es imposible analizar manualmente todos los sitios web en busca de posibles plagios. Por este motivo, las empresas se han visto en la necesidad de buscar métodos automáticos capaces de desempeñar dicha labor. Estos métodos deberán señalar los documentos sospechosos a un profesional de la propiedad intelectual el cual decidirá si se trata de un caso de plagio, puesto que el plagio es un juicio académico que no puede ser medido por una máquina [30]. De esta forma, el número de sitios a analizar disminuye drásticamente y la empresa se siente más protegida ante la competencia desleal.

1.2.1. Plagio de ideas

Como ya hemos comentado, Internet se ha convertido en el medio de comunicación más utilizado por las empresas para acercarse al consumidor. Las empresas crean páginas web donde introducen información propia de la empresa, publicitan sus productos y sus servicios, sus nuevas herramientas, sus ideas originales, sus logros, sus metas, sus principios, etc... Con estas páginas las empresas pretenden estar más en contacto con los consumidores y conseguir que los usuarios descubran los productos ofertados y las novedades ofrecidas.

Sin embargo, la información en Internet es visible para todos, no sólo para los consumidores sino también para las empresas competidoras. Cuando una empresa lanza una herramienta nueva, introduce una funcionalidad original, cambia el formato de la web, tanto consumidores como competidores lo descubren en pocas horas o días.

Si una empresa quiere estar en primera línea de salida, debe de estar atento a sus competidores, es decir, comparar los productos de unos y otros, descubrir las novedades, el efecto que tienen estas novedades en los consumidores. De esta forma poder mejorar los productos o herramientas que ofrece el resto de empresas. Pero no todas las empresas son leales, sino que existen empresas que utilizan la información que introducen otras empresas en sus páginas web para copiar las ideas de éstas.

Un ejemplo (véase la figura 5.1) lo protagoniza la empresa "MVS Technologies", la cual plagia en su total extensión y letra por letra las ideas y servicios que ofrece la empresa "Asesoria Informática"[5].

Otras empresas, en cambio, copian las páginas de otras empresas modificando el contenido, incluso llegando a inventarse instituciones que no existen. Un ejemplo es la empresa de Lima "Peru Creative", que copió los casos de éxito de la empresa valenciana "Adding Technology", la cual colabora con la Generalitat Valenciana (véase la figura 1.2). El resultado del plagio fue bastante penoso, ya que se inventó una Generalitat Limana, y por si fuera poco se olvidó de cambiar el logotipo de la Generalitat Valenciana[6].

Incluso hay empresas que copian las campañas de publicidad. Por ejemplo, la empresa

[5]http://www.consultoriainformatica.net/

[6]http://www.levante-emv.com/ (Levante-EMV, 2007)

Figura 1.1: Ejemplo de plagio de ideas

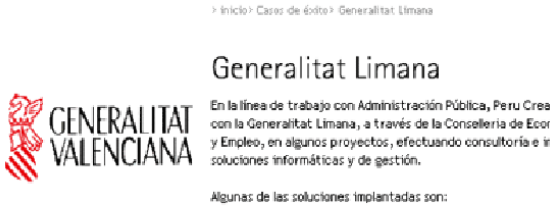

Figura 1.2: Ejemplo de plagio de casos de éxitos

"Corbeta Marketing y Compañía" (véase la figura 1.3) plagió textualmente la campaña en Google, haciendo uso incluso del mismo slogan *Navegue con Expertos* de la empresa "WEB-SEO.CL". En este caso se descubrió el plagio porque aparecían las dos empresas anunciadas una detrás de la otra con el mismo texto y agregándole solamente su dirección web, hecho que se ha prestado para que los clientes de la empresa plagiada piensen que esta empresa era su filial[7].

Hay incluso empresas que plagian información correspondiente a los servicios que ofrecen otras. Por ejemplo, la empresa "Buquebus" demandó a "Pluna" por plagiar su proyecto de vuelos regionales[8].

Estos son un pequeño ejemplo de la gran cantidad de ataques a la propiedad intelectual. Es por esto que las empresas se sienten indefensas ante el ataque a sus ideas y sus innovaciones. Por eso las empresas deben de protegerse ante cualquier intento de plagio por parte de competidoras. En este trabajo, mostraremos diferentes herramientas que una empresa puede utilizar para protegerse del plagio.

[7]http://lacorbeta.wordpress.com/2010/03/29/plagio-a-nuestra-campana-navegue-con-expertos/
[8]http://www.miradornacional.com/(Mirador Nacional, 2009)

Figura 1.3: Ejemplo de plagio de campañas de publicidad

1.2.2. Plagio de opiniones

El plagio no solamente afecta a las empresas sino también a los consumidores. Las opiniones de éstos se ven gravemente afectadas con casos de plagio. Y, como ahora cualquiera puede publicar de forma gratuita y sencilla, el plagio empieza a alcanzar proporciones verdaderamente alarmantes.

En ocasiones alguien publica alguna nota en un blog como *slashdot.com*, posteriormente otro la copia para publicarla en *barrapunto.com*. Otro tanto ocurre en las blogs particulares; por ejemplo, alguien publica alguna opinión en su blog particular y posteriormente otro *bloguero* la publica en su blog también particular sin introducir ninguna referencia a la opinión original. Casos como estos son muy frecuentes en el mundo de las redes sociales.

Una de las principales causas es que las redes sociales miden su éxito en función del número de páginas visitadas o de la cantidad de amigos que se genere. Es decir, que el deseo de fabricar contenidos para atraer la atención de la gente se ha vuelto tan fuerte, que la tentación de copiar se ha vuelto irresistible.

A esto hay que sumarle que el relativo anonimato que nos proporcionan los blogs nos hace sentir seguros, ya que puede verse como un forma de protección para no ser descubierto. Además, esto puede conllevar un beneficio económico, puesto que cuantas más visitas se consiguen, mayores serán los beneficios por publicidad.

1.2.3. Prevención de pérdida de datos

Toda empresa corre un constante peligro debido a la existencia de personas ajenas a la información, conocidas como piratas informáticos o *hackers*, que buscan tener acceso a la red empresarial para modificar, copiar o borrar datos. Los administradores de sistemas están horas, e incluso varios días, volviendo a cargar o reconfigurando sistemas comprometidos, con

el objetivo de recuperar la confianza en la integridad del sistema [79].

El resultado de una pérdida de datos es la atención negativa de los medios, la reducción de la confianza de los clientes y socios, una reducción de valor de la empresa, el daño a la reputación, pérdida de competitividad y posibles cargos criminales. Proteger los datos sensibles es crucial para la mayoría de las organizaciones donde la propiedad intelectual y la información confidencial está relacionada con el valor monetario de la empresa.

Tomar conciencia del valor de la información corporativa fue el objetivo de la IV Jornada Internacional de ISMS Forum Spain[9] (asociación española para el fomento de la seguridad de la información) en Sevilla, donde se reunieron los CIO (*Chief Information Officer* o líder de las tecnologías de la información) de empresas españolas, que debatieron junto a expertos internaciones los últimos desarrollos en tecnologías DLP (*Data Leak Prevention* o Prevención de Pérdidas de datos). En esta Jornada, Rustem Khayretdinov, vicepresidente de ventas de Infowatch[10] y un firme convencido que en el futuro todo el mercado de la seguridad estará enfocado a la protección de datos, comentó:

> *...la pérdida de la información significa un lastre para la competitividad... Toda política de seguridad y protección de datos debe comenzar por entender y determinar la información que se debe proteger, de qué manera y de quién. A partir de este punto se debe decidir los cambios en los procesos de negocio para gestionarlos adecuadamente.*

Y es que la información que posee una empresa es uno de los principales activos a proteger. Se han propuesto varias técnicas para proteger la información de ataques externos. Una de estas técnicas es utilizar los métodos para detección automática de plagio para prevenir estos ataques a la red informática y así poder evitar la pérdida de datos. Estos trabajos tratan a las entradas al sistema como datos secuenciales. Las secuencias de símbolos discretos son una de las representaciones de datos fundamentales en informática. Una gran cantidad de aplicaciones dependen de análisis de datos secuenciales, desde motores de búsqueda, hasta las herramientas de vigilancia de redes y los programas antivirus [96]. De esta forma pueden encontrar patrones en las cargas útiles de paquetes de red de las entradas al sistema, y así pueden identificar aquellas entradas sospechosas de ser ataques al sistema.

1.3. Análisis de opiniones

Para una empresa tanto la cantidad como la calidad de información no tiene precio. El principal objetivo de la información es la de apoyar a la toma de decisiones, puesto que con ella se tendrán más bases sustentables para poder decidir cuáles son los pasos a seguir y qué rumbo hay que tomar para lograr los objetivos que se planificaron; es más, gracias a la información, se contarán con un mayor número de armas para afrontar el camino que decidirá el

[9] http://www.ismsforum.es/
[10] http://www.infowatch.com/

futuro de la organización. Es por ello que en una empresa se le debe de poner una atención sumamente especial a la información que se genera cada día, la adecuada interpretación de ésta establecerá los cimientos necesarios para consolidarse como una empresa de éxito en el mercado y se obtendrá una mayor oportunidad de crecimiento y expansión de mercado. La información le permitirá identificar cuáles son las fortalezas con las que cuenta y cuáles las debilidades y sectores vulnerables que presenta como organización. Teniendo en cuenta estos datos podrá tener una planificación más alcanzable y factible, ya que podrá identificar donde tiene que aumentar los esfuerzos y qué parte de la empresa necesita mayor atención [13].

Asó como comentamos en el apartado anterior una de las clases de información más importante para una empresa es la opinión de los consumidores, las cuales marcan las tendencias del mercado. Con este propósito muchas empresas gastan enormes cantidades de dinero en encuestas de satisfacción del cliente, para obtener sus opiniones sobre los productos o servicios ofertados. Pero en muchos casos estas encuestas no son eficaces, tanto por la dificultad de conseguir un número elevado de encuestados como por la dificultad de conseguir encuestas eficaces [84]. Por este motivo, las empresas tienen la obligación de encontrar otros medios para conseguir dicha información.

Actualmente los avances tecnológicos han generado un ritmo acelerado de cambio en el marketing, tanto en la oferta de productos como en los canales de comunicación [45]. Y es que las empresas han visto en las redes sociales una oportunidad inmejorable para obtener información de primera mano de los consumidores y, por tanto, un medio en el que comienza a centrarse el marketing de las empresas. Este hecho se ve reflejado en que investigadores de las ciencias empresariales están comenzado a centrarse en las redes sociales digitales como tema de investigación actual, e incluso se comienzan a celebrar congresos enfocados a esta área, como por ejemplo el *Global Marketing Conference 2010 (Tokio, Japón)*[11] con ponencias especializadas en publicidad interactiva, en el fenómeno de "boca a oreja" (*Word of Mouth*), en los contenidos generados por los usuarios y *Mobile Marketing* [98]. Algunos de estos estudios realizados en el ámbito de las redes sociales digitales y su importancia para el marketing de las empresas son:

- El trabajo de Harris y Rae [38] busca determinar el poder de las redes sociales digitales en la construcción de reputación de marca y relaciones con los clientes, en pequeñas y medianas empresas.

- Zhang y Watts [129] investigan en qué medida el concepto de comunidades de práctica se puede aplicar a las comunidades online y explora cómo las organizaciones pueden utilizar mejor las estructuras sociales digitales para la práctica de gestión del conocimiento.

- En el trabajo de Jansen et al. [50] se investiga el *microblogging* como una forma de "boca a oreja" para el intercambio de opiniones de los consumidores con respecto a las marcas.

[11] http://www.kamsconference.org/

Y es que analizando las compras y los servicios contratados que se efectúan en la actualidad, se observa que una gran mayoría de éstos no están condicionados por las sugerencias de las campañas de publicidad y los trucos del marketing, sino por los comentarios que otros consumidores han escrito en los múltiples foros virtuales (públicos y privados) que hoy ofrece la Web 2.0 [81]. Con la explosión de la Web 2.0, plataformas como blogs, foros de discusión, redes sociales, y varios otros tipos de medios de comunicación sociales, los consumidores tienen a su disposición un lugar donde compartir sus experiencias con las diferentes marcas y donde poder dar sus opiniones, positivas o negativas sobre cualquier producto o servicio. Las empresas principales empiezan a darse cuenta que estos comentarios de los consumidores pueden manejar la enorme influencia en la formación de las opiniones de otros consumidores y, en última instancia, su lealtad a la marca, sus decisiones de compra.

Las empresas que deseen crecer (o llegado el caso, sobrevivir) deben de responder a las perspicacias de los consumidores, y es por todo esto que tienen la obligación de analizar los medios de comunicación sociales, para obtener la información adecuada para modificar sus mensajes de marketing, modificar el desarrollo de los productos, mejorar los servicios, etc [127] y de este modo acercarse al consumidor. Las redes sociales jugarán un rol clave en el futuro del ejercicio del marketing, porque pueden ayudar a reemplazar el disgusto del cliente por la fidelización. Las empresas que prosperarán serán las que de forma proactiva se identifiquen y hagan uso de este nuevo mundo, porque consideran el cambio como una oportunidad más que como una amenaza que hay que evitar a toda costa [38].

Y es que gracias a la Web 2.0 gana peso la opinión del ciudadano frente a las marcas y sus técnicas comerciales más tradicionales. Según un estudio, realizado por la compañía de software Six Apart[12], el 75 % de los encuestados reconocían que sus decisiones de compra están directamente influenciadas por lo que leen en blogs, foros y el resto de medios de comunicación sociales [81]. Según el Instituto Nacional de Estadística (INE[13]), el 80 % de los internautas reconoce que acude a la red para informarse sobre productos, marcas y servicios. En otro estudio reciente realizado en febrero de 2009 por la Asociación para la Investigación de Medios de Comunicación (AIMC[14]), el 75.5 % de internautas españoles admite haberse documentado en internet durante el último año, como paso previo a formalizar una compra de productos o servicios, bien sea para realizar dicha adquisición de manera online o de manera offline.

Ante esta situación las empresas se preguntan qué es lo que lleva a los consumidores a fiarse de la opinión de un tercero al cual no conocen y desconfían de la marca que les acompañó toda la vida. Los sociólogos Víctor Gil y Felipe Moreno en [33] responden a este interrogante aduciendo a la competencia entre las empresas y al colaborismo entre los consumidores:

> *La clave está en que, mientras las marcas compiten entre sí, los consumidores colaboramos entre nosotros. La Web 2.0 invita a una nueva forma de comunicación, basada en la conversación y la cooperación, que potencia la honestidad y*

[12] http://www.sixapart.com
[13] http://www.ine.es
[14] http://www.aimc.es/aimc.php

> *la transparencia. Son otros códigos y otro nivel de confianza... Nos hemos convertido en consumidores sofisticados... Tantos años de publicidad han terminado convirtiéndonos en expertos en marketing y, de paso, en unos desconfiados ante las marcas.*

Dadas estas circunstancias, los responsable del marketing de las empresas tienen la obligación de supervisar en los medios de comunicación sociales las opiniones relacionadas con sus productos y servicios, e incluso las opiniones de los consumidores de los productos de las empresas competidoras. Sin embargo, en lo últimos años se ha producido una explosión en la Web 2.0 sin precedentes, ocasionando que la supervisión manual de las opiniones de los consumidores se convierta en un trabajo completamente irrealizable. La empresa especializada en blogs Technorati[15] estima que 75,000 nuevos blogs son creados diariamente, con 1.2 millones de nuevos comentarios cada día en las que el consumidor comenta sobre productos y servicios [53]. Por este motivo, en los últimos años están tomando protagonismo aplicaciones de búsqueda de opiniones sobre productos, marcas y personas, como son: Social Mention[16], Same Point [17], Flock [18] y User Voice [19].

Aunque estas aplicaciones son capaces de encontrar las opiniones generadas por los consumidores en las redes sociales sobre productos y marcas, las empresas se ven en la necesidad de aunar esfuerzos por encontrar un método automático que sea capaz de analizar dichas opiniones e identificar su orientación semántica. Es decir, un método para identificar si las opiniones de los consumidores sobre los productos o servicios son positivas o negativas [115].

1.3.1. Basado en ontologías

Sin embargo, una opinión generalmente positiva sobre algún producto concreto, no significa que el titular de la opinión exponga opiniones positivas sobre todos los aspectos o características del objeto. Del mismo modo, una opinión negativa no significa que el opinante no esté conforme con todos los aspectos del producto o servicio. En un documento de opinión, tales como una opinión del cliente sobre un producto o servicio, el titular de la opinión describe tanto aspectos positivos como negativos del objeto, aunque el sentimiento general del objeto puede ser positivo o negativo.

Como se ha comentado anteriormente, las empresas están interesadas en conocer la opinión general de un producto o servicio ofrecido; no obstante, también están interesadas en conocer qué conceptos pueden mejorarse o reforzarse. Por ejemplo, una empresa de turismo que ofrece un viaje a Milán, con el hotel *Noches en Milán* incluido, y entradas a una ópera en la *Scala di Milano*; es lógico pensar que la empresa estará interesada en conocer si el viaje es del gusto de los clientes. Analizando las opiniones de los clientes, aparecen opiniones como:

[15] http://www.technorati.com/
[16] http://www.socialmention.com
[17] http://www.samepoint.com/
[18] http://flock.com/
[19] https://uservoice.com/

El hotel "Noches en Milán" no nos ha gustado nada, era desastroso y muy ruidoso, no lo recomiendo a nadie; pero la ópera en la Scala di Milano era una maravilla

En esta opinión, que puede calificarse como una opinión generalmente negativa, aparecen dos polaridades diferentes: el concepto *hotel* tiene una polaridad negativa; pero por otro lado, el concepto *ópera* tiene una polaridad positiva. Si la empresa sólo analiza la orientación semántica general de la opinión, pierde la información de que al opinante le ha gustado las entradas ofrecidas para la ópera. En el caso que la mayoría tengan la misma opinión, la empresa podría dejar de ofrecer el viaje a Milán. Sin embargo, esta empresa perdería una oportunidad de negocio, puesto que estudiando las orientaciones semánticas de los conceptos sobre los que se opinan, podría descubrir que lo que no gusta a los clientes es el hotel y no el viaje. Tal vez, cambiando de hotel ofrecido en el viaje, mejore las opiniones de los clientes sobre el viaje.

Por tanto, analizar no sólo la polaridad general de una opinión es importante, sino que también lo es analizar la polaridad de cada concepto calificado en las opiniones. Esto ayudará a una empresa a mejorar el producto o servicio según los gustos de los clientes.

Para poder analizar la polaridad de los conceptos que se opinan en los documentos evaluativos, las empresas pueden aprovecharse de las ontologías que poseen. Las empresas disponen de ontologías en las que están representados todos los aspectos de los productos y servicios que ofrece. A partir de las ontologías se facilitaría la extracción de las opiniones sobre cada concepto.

Volviendo al ejemplo anterior, si la empresa de turismo posee una ontología con un concepto *hotel* y otro concepto *ópera*, podría extraer los adjetivos asociados a cada concepto y a partir de éstos calcular la polaridad de cada uno de los conceptos.

1.3.2. Vía fusión de ontologías

La capacidad de los humanos para tener diferentes puntos de vista de un mismo concepto no tiene límites. Por ello, no es de extrañar que diferentes analistas, desarrolladores e ingenieros del conocimiento tengan diferentes conceptos de la realidad sobre un mismo dominio. Esta es la principal causa de que un dominio pueda tener múltiples ontologías completamente diferentes. Unas estarán más detalladas, otras estarán enfocadas a una parte concreta del dominio, etc... Por tanto, es lógico deducir que dos empresas dedicadas al mismo dominio posean diferentes ontologías.

Sin embargo, dado el coste de conseguir la opinión de los consumidores, puede que varias empresas decidan compartir e intercambiar la información que poseen sobre las opiniones de los consumidores, o incluso, llegado el caso extremo en el que dos empresas se fusionen, no se quiere perder ningún dato de los análisis de opiniones que han realizado cada una de las empresas con anterioridad. En estos casos, se debe de encontrar algún método que sea capaz de poder analizar automáticamente las opiniones de los clientes y además que sea compatible con las diferentes ontologías.

Esta posibilidad de intercambio de información de opiniones no se ha estudiado anteriormente. En este trabajo proponemos un algoritmo que incluye dentro del análisis de opiniones, una fusión de ontologías. La fusión de ontologías nos facilitará poder obtener las polaridades de cada concepto de cada una de las ontologías de las empresas participantes. Esto es posible ya que la fusión de ontologías nos devolverá una alineación entre cada concepto de las dos ontologías de las empresas con lo que podremos relacionarlos y así obtener la polaridad de dichos conceptos.

Para comprobar la eficacia del algoritmo hemos realizado un estudio sobre el análisis de opiniones vía fusión de ontologías dentro del dominio del turismo. Para ello nos ponemos en la piel de dos empresas dedicadas al turismo las cuales analizarán opiniones de consumidores sobre conceptos de dicho dominio. Posteriormente, realizaremos una simulación en la que dichas empresas deciden compartir la información.

1.4. Estructura de la tesis

Este documento está organizado del siguiente modo:

- En el Capítulo 2 exponemos el estado del arte del campo de la detección automática del plagio. Además expondremos algunas de las herramientas disponibles que las empresas pueden utilizar para protegerse contra las infracciones de la propiedad intelectual.
- En el Capítulo 3 describimos la competición PAN'09 en la cual hemos participado para comprobar la eficacia de la herramienta WCopyFind.
- En el Capítulo 4 introducimos el estado del arte de la minería de opiniones. Debido a que en nuestro algoritmo introducimos una fase de fusión de ontologías, también se hará una introducción al estado del arte de dicha disciplina.
- En el Capítulo 5 describimos los experimentos que hemos realizado para la evaluación del algoritmo propuesto. En este capítulo también describimos los experimentos que hemos realizado para la selección de un método de fusión de ontologías eficaz.
- En el Capítulo 6 exponemos las conclusiones a las que hemos llegado en este trabajo y las líneas a seguir en un futuro.

Capítulo 2

Detección automática de plagio

Tras el cambio que ha producido en la sociedad los medios digitales, las empresas cada vez son más vulnerables a ser víctimas de casos de plagio. Esto es debido por dos razones principales: porque existe la falsa idea de que se puede utilizar con total libertad el material que se publica en la Web por el hecho de ser de dominio público [79] y, porque debido a la enorme cantidad de información en la Web se puede llegar a creer que existe una probabilidad muy pequeña de que pueda ser descubierto.

Lo que diferencia una empresa de otra son sus ideas, sus productos, sus éxitos anteriores, etc. Si una empresa competidora le plagia tanto ideas como productos, la empresa plagiada pierde aquello que la diferencia del resto, por lo que pierde a su vez la ventaja que le proporcionarán dichas ideas. Es por eso que las empresas están obligadas a protegerse contra cualquier ataque a sus propiedades intelectuales.

En este capítulo repasaremos el estado del arte de la disciplina de la detección automática de plagio. Además introduciremos algunas de las herramientas disponibles que una empresa puede utilizar para detección automática de plagio.

2.1. Estado del arte

Los diferentes métodos que existen actualmente en la disciplina de detección automática de plagio se pueden agrupar en tres diferentes grupos: análisis intrínseco de plagio, detección del plagio con referencia y detección del plagio translingüe. Además de estos tres grupos existen algunos investigadores que han visto interesante utilizar la detección automática de plagio para usos de prevención de pérdida de datos. Más concretamente para ataques por parte de intrusos en redes informáticas. En esta sección introduciremos cada uno de los tres grupos de métodos, así como las investigaciones realizadas para prevención de pérdidas de datos en sistemas informáticos.

2.1.1. Análisis intrínseco del plagio

La mayor dificultad que encontramos cuando queremos descubrir si un documento es plagio de algún otro, es poder comparar dicho documento sospechoso con todos los documentos disponibles. Es más, podemos afirmar que dicha comparación es una tarea imposible de realizar. Además, en ocasiones no se tiene un conjunto de documentos sospechosos contra los cuales comparar. Por eso, algunos investigadores han comenzado a estudiar métodos automáticos de detectar el plagio sin referencia. Este tipo de estudios se llama *análisis intrínseco del plagio*.

En [108], Stein y Meyer zu Eissen definen el problema del análisis intrínseco del plagio de la siguiente manera:

> *Dado un documento d, supuestamente escrito por un autor, y queremos identificar los fragmentos en d que se derivan de otro autor y que no estén citados adecuadamente. El análisis intrínseco del plagio es un problema de clasificación de una sola clase. La propiedad del problema más destacada de esta clasificación es que la información de una sola clase está disponible. Esta clase se llama la clase de destino, todos los demás objetos están comprendidos en la clase de los llamados valores atípicos. En el contexto del análisis intrínseco del plagio todos los documentos, o partes del documento del pretendido autor, forman la clase de destino, y todos los documentos, o partes del documento de otro autor arbitrario, constituyen la clase atípica. Hay que destacar que el documento d es la única fuente para formular un modelo de estilo de escritura para los objetos de la clase objetivo, mientras que la formulación de este modelo se ve dificultado en la medida en que d es un plagio. También hay que destacar que los documentos en la clase de valores atípicos son tan abundantes que ni una muestra representativa, ni la formulación de un modelo de estilo de escritura de esta clase es posible.*

El análisis intrínseco del plagio es una disciplina joven que fue introducida por los mismos autores, Stein y Meyer zu Eissen en [107]. Los principios básicos del análisis intrínseco del plagio son [16]:

- cada autor tiene su propio estilo de escritura;
- el estilo de escritura de cada autor debería seguir siendo coherente en todo el texto;
- las características de un estilo es difícil de manipular y de imitar, haciendo destacar el fragmento del trabajo plagiado.

El estilo de escritura de un autor incluye diferentes características como la riqueza en el vocabulario, la longitud oracional o el número de palabras de paro. El análisis de estas medidas permite a los investigadores plasmar el estilo de escritura en números reales [11]. Aunque el análisis intrínseco de plagio sea una disciplina joven, cuantificar el estilo de escritura es

una línea de investigación abierta desde la década de 1940 [126, 29]. Varios métodos han sido propuestos para medir las características del estilo de escritura, como:

- Métodos para medir la riqueza de vocabulario:
  - En [44], Honore propuso la función R, la cual mide la variedad de vocabulario en un texto. La función R se define como:

$$R = \frac{100 \log N}{1 - (V_1/V)} \tag{2.1}$$

    donde N es la longitud del texto, V es el número de diferentes palabras del texto y V_1 es el número de diferentes palabras que aparecen una sola vez en el texto. Está función nos indica que, cuantas más palabras tenga el texto que no se repiten, mayor riqueza de vocabulario tendrá el texto. Es más, si todas las palabras aparecieran una sola vez (es decir, si $V_1 = V$), la función R tendería al infinito [43].
  - En [126], Yule ideó la función K, una medida de la riqueza de vocabulario basado en el supuesto de que la aparición de una palabra dada se basa en el azar y puede considerarse como una distribución de Poisson.

$$K = \frac{10^4(\sum_{i=1}^{\infty} i^2 V_i - N)}{N^2} \tag{2.2}$$

    donde N es la longitud del texto y V_i es el número de diferentes palabras que aparecen i veces en el texto.
  - En [102], Sichel propuso un modelo teórico para las distribuciones de vocabulario. La distribución de Sichel calcula cual es la probabilidad de que cualquier tipo de palabra aparezca exactamente r veces en un texto de longitud N, y se define como:

$$\theta(r|N) = \frac{2\alpha/\pi)^{1/2} \exp \alpha}{\exp[\alpha\{1 - (1-\theta)^{1/2}\}] - 1} \frac{(1/2\alpha\theta)^r}{r!} K_{r-1/2}(\alpha) \quad (r = 1, 2, ..., \infty) \tag{2.3}$$

    donde $\alpha > 0$, $0 < \theta < 1$ y $K_{r-1/2}(\alpha)$ es la función de Bessel [63] modificada de segunda especie de orden $r - 1/2$ y argumento α.
- Métodos para medir la complejidad y comprensibilidad del texto:
  - En [29], Flesch propone la Prueba de Legibilidad de Flesch (*Flesch Reading Ease*), para medir la complejidad de un documento. Se define como:

$$RE_F = 206385 - 1015\frac{N}{F} - 84.6\frac{S}{N} \tag{2.4}$$

    donde N es el número de palabras del documento, F es el número de frases y S el número total de sílabas. *Flesch Reading Ease* otorga al texto un valor real, el cual cuanto más elevado sea el resultado, más fácil será comprender el documento. Por ejemplo, un resultado entre 60 y 70 el documento es fácilmente comprensible para personas mayores de 15 años.

- En [106], Meyer zu Eissen y Stein propusieron el cálculo del promedio de clases de palabras basado en la frecuencia. En este método cada palabra $w \in D$ es asignada a una clase $C(w)$, la cual se asigna según la frecuencia de la palabra $f(w)$, de la siguiente manera:

$$C(w) = \lfloor \log_2(f(w^*)/f(w)) \rfloor \tag{2.5}$$

  donde w^* denota la palabra más frecuente en D. La clase asociada a w^* es C_0.

- Métodos para determinar los requisitos específicos del lector que son necesarios para comprender un texto, como la graduación:

  - En [55], Kincaid y sus colaboradores presentan *Flesch-Kincaid Grade Level* que es una mejora de la fórmula de Flesch. La fórmula se define como:

$$GL_{FK} = 0.39\frac{N}{F} + 11.8\frac{S}{N} - 15.59 \tag{2.6}$$

    donde N es el número de palabras del documento, F es el número de frases y S el número total de sílabas. El resultado es un número que se corresponde con un nivel de graduado.

  - En [14], Dale y Chall proponen la fórmula de legibilidad de Dale-Chall (*Dale-Chall Readability Formula*). Dicha fórmula se expresa como:

$$RF_{DC} = 0.1579\alpha + 0.0496\beta + 3.6365 \tag{2.7}$$

    donde α es el porcentaje de palabras difíciles y β es la media de la longitud de las frases.

Para comprobar que un texto o un fragmento de texto es plagio, no es suficiente calcular dichas características para todo el documento. En [107] proponen el cálculo de diversas características considerando el documento sospechoso completo, y posteriormente proponen realizar los mismos cálculos para cada fragmento del documento. Una vez realizados todos los cálculos, comparan los resultados para comprobar si sufren grandes variaciones, las cuales indicarán que dichos fragmentos son candidatos a ser plagio.

En [109] comprueban la robustez de las medidas de la riqueza de vocabulario, tales como la función K de Yule, la función R de Honore, y el promedio de clases de palabras basado en la frecuencia. Los resultados demostraron que únicamente se puede afirmar que es estable el promedio de clases de palabras basado en la frecuencia, ya que proporciona resultados fiables incluso en fragmentos cortos.

En [105] el autor propone la utilización de una función de cambio de estilo basado en una apropiada medida de disimilitud originalmente propuesta para la identificación del autor. Además, proponen un conjunto de reglas heurísticas que intentan detectar los documentos plagiados libremente y pasajes plagiados, así como para reducir el efecto de cambios irrelevantes de estilo en los documentos. Con este método, Stamatatos, ganó la primera edición de

la competición de plagio (*1st International Competition on Plagiarism Detection*[1] (PAN'09)), en la tarea de análisis intrínseco de plagio.

Sin embargo, debemos hacer notar que el análisis intrínseco del plagio no demuestran que los fragmentos son plagios, puesto que este análisis es incapaz de indicar los textos originales. Esto se debe principalmente a que en los cálculos no se hace ningún tipo de comparación con textos originales [5]. El objetivo del análisis intrínseco del plagio es indicar algún fragmento del texto en el que se produzca un cambio de estilo de escritura, característica que podría implicar un caso de plagio.

2.1.2. Detección del plagio con referencia

La detección del plagio con referencia se basa en la comparación de un conjunto de documentos sospechosos de ser plagiados con un conjunto de documentos originales. La detección del plagio con referencia busca fragmentos que pueden estar duplicados en otros documentos. Estos fragmentos pueden ser párrafos, un conjunto de frases o un conjunto de palabras. Una vez se localiza los posibles fragmentos duplicados será un humano quien decida si los fragmentos son realmente fragmentos plagiados.

Una posible solución a este problema es comparar cada pasaje de un documento sospechoso con cada pasaje de cada documento en el corpus de referencia. Los tiempos de ejecución serán obviamente prohibitivos [83]. Además a medida que aumentamos el corpus de referencia mayor será el tiempo de ejecución. Existen estudios que intentan reducir el espacio de búsqueda, como en [7], donde los autores proponen la reducción del espacio de búsqueda en base a la distancia de Kullback-Leibler.

Hasta la actualidad existen una gran variedad de diferentes métodos para la detección del plagio con referencia. Existen diferentes puntos de vista para decidir que unidad de comparación es la más adecuada. Para decidir cual es la mejor unidad de comparación, hay que tener en cuenta diferentes factores, como que cuanto más grande sea la unidad de comparación menor es la probabilidad de señalar fragmentos sospechosos de documentos no relacionados. Por ejemplo, dos documentos independientes pueden ambos tener una frase como "en esta investigación el autor nos presenta" como parte de un párrafo. Si la unidad de comparación es un párrafo, probablemente no se detectarán como un fragmento sospechoso de plagio, mientras que si se detectaría si la unidad de comparación es una frase, cuando en realidad no lo es. Por tanto, cuanto mayor es la unidad de comparación, la *precisión* aumentará.

Por otro lado, también debemos tener en cuenta que cuanto mayor sea la unidad de comparación mayor es la probabilidad de no detectar documentos realmente plagiados. Por ejemplo, consideremos dos documentos que comparten 5 o 6 frases. Si la unidad de comparación es un párrafo, probablemente no se detectarán como un fragmento sospechoso de plagio cuando en realidad sí lo es, mientras que sí se detectaría si la unidad de comparación es una

[1] http://pan.webis.de/

frase. Por tanto, cuanto mayor es la unidad de comparación, la *cobertura* disminuirá.

Con independencia de la unidad de comparación utilizada, todos los métodos de detección de plagio necesitan una medida de similitud para comparar los fragmentos de textos correspondientes a la unidad de comparación. La mayoría de las medidas de similitud utilizadas en la detección de plagio se basan en la estimación de la cantidad de configuraciones comunes de las palabras. Se diferencian por las sub-cadenas que lo forman (n-gramas, subsecuencias, etc) o por las palabras utilizadas en las comparaciones (solamente las palabras que forman los fragmentos de texto, o incluyendo sinónimos de WordNet, etc.) [37].

En [69], Lyon *et al.* proponen el uso de trigramas para medir la similitud entre los textos. Los autores basan su elección en el hecho de que el número de trigramas comunes en dos textos independientes debería ser bajo (incluso si los textos pertenecen al mismo tema), debido a la distribución Zipf [132] de las palabras. Barron y Rosso en [6] coinciden con Lyon y sus colaboradores. En este estudio, los autores realizaron una comparación entre la frecuencia de n-gramas comunes entre diferentes documentos demostrando que la probabilidad de encontrar n-gramas comunes en diferentes documentos decrece conforme se incrementa n. La tabla 2.1 muestra los resultados obtenidos en este estudio. Además, Barron y Rosso muestran que bigramas y trigramas son las mejores unidades de comparación para la detección automática de plagio. Finalmente concluyeron que con bigramas se mejora el *recall* y con trigramas se obtiene mejor *precisión*.

Documentos	2-gramas	3-gramas	4-gramas
2	0.1125	0.0574	0.0312
3	0.0302	0.0093	0.0027
4	0.0166	0.0031	0.0004

Tabla 2.1: n-gramas comunes en diferentes documentos

También existen trabajos que no utilizan los n-gramas como unidad de medida de comparación. Por ejemplo, Shivakumar y Garcia-Molina en [101] propusieron SCAM uno de los primeros trabajos enfocados a nivel de documento. Los autores argumentan que si dos documentos tienen frecuencias similares de ocurrencia de un grupo de palabras, es probable que se traten de diferentes versiones de un mismo documento [5].

Kang *et al.* en [52] presentaron PPChecker, el cual está basado en una medida de similitud que tiene en cuenta los sinónimos de las palabras de los documentos sospechosos, obtenidos a través de WordNet. Esta solución intenta corregir el cambio de redacción producido cuando el autor de plagio modifica las palabras de los documentos originales por sinónimos. Para Kang y sus colaboradores el proceso de comparación debe hacerse a nivel de sentencia [5].

2.1.3. Detección del plagio translingüe

Algo muy común cuando se comete plagio, es la traducción de un idioma a otro idioma sin indicar como referencia el documento origen. Cuando esto ocurre, los sistemas mencionados anteriormente no son capaces de detectarlo. Para detectar automáticamente este tipo de plagio se debe tener otro enfoque diferente.

Probablemente los primeros trabajos dedicados al minado de texto translingüe sean el estudio de Resnik [89] y el de Chen y Nie [49]. Ambos trabajos realizaban primero un sistema de minería de Internet para extraer textos paralelos en diferentes idiomas. Estos sistemas se basaban en motores de búsqueda convencionales para encontrar páginas en las que aparecía el texto "versión en Inglés" y "versión en Francesa" en etiquetas de hiper-vínculos en HTML. Como resultado consiguieron, para cada sitio web, un pequeño número de diferentes traducciones. Aunque estos documentos no se podían considerar como plagio, sí que fueron útiles en la fase de entrenamiento.

Resnik [89] aplicó una herramienta de identificación de idiomas y utilizó características estructurales para alinear los dos textos. Posteriormente, para identificar si un texto era traducción de otro utilizó el criterio de longitud de la cadena de los fragmentos alineados, es decir, un fragmento alineado era traducción de otro si tenían una longitud similar.

Chen y Nie [49] utilizaban las características de texto como la URL, el tamaño, la fecha, el idioma y un conjunto de caracteres para determinar si los documentos eran traducciones de otro. Con este sistema obtuvieron en sus experimentos una precisión del 95 % para pares de documentos de Inglés-Francés y del 90 % para Inglés-Chino.

Pouliquen y sus colaboradores [94] diseñaron un sistema de trabajo que podía identificar a las traducciones y otros documentos muy similares entre un gran número de candidatos, al representar el contenido de los documentos con un vector de términos a partir del tesauro Eurovoc[2], para posteriormente medir la similitud semántica entre los vectores. En las pruebas realizadas en el estudio, obtuvieron una precisión del 96 % en un espacio de búsqueda de 820 documentos.

En [92] los autores proponen un método basado en tres fases principales. La primera fase consiste en dado un documento sospechosos d y un corpus referencia C en diferentes idiomas, recuperar un subconjuntos de documentos candidatos de C, los cuales podrían ser fuentes de fragmentos plagiados del documento d. El siguiente paso, se realiza un análisis semántico entre las secciones de d y cada documento $c_i \in C$. Por último, se analizan las secciones similares buscando alguna cita adecuada, lo cual los eliminaría del posibles plagios.

Un estudio más reciente de Barron et. al [8] se basó en el uso de un diccionario estadístico-bilingüe basado en el modelo IBM-1. Este diccionario se creó a partir de un corpus paralelo que contenía fragmentos originales escritos en un idioma y versiones plagiadas de estos fragmentos escritos en otro idioma. El objetivo del trabajo era crear un sistema capaz de detectar el plagio multilingüe con respecto a un autor determinado.

[2] http://europa.eu/eurovoc/

En [128] Ceska y sus colaboradores describen MLPlag, el cual es un método para la detección de plagio en un entorno multilingüe. Este método se basa en el análisis de las posiciones de palabra. Utiliza el diccionario de sinónimos EuroWordNet[3] que transforma las palabras de forma independiente del lenguaje. Esto permite identificar los documentos plagiados a partir de fuentes escritas en otras lenguas. También se incorporaron técnicas para identificar la sustitución de palabras por sinónimos.

En la primera edición de la competición de detección automática del PAN realizado en 2009, ninguno de los equipos participantes consiguió detectar casos de plagio translingüe. Esto es un claro ejemplo de la dificultad que conlleva esta tarea.

2.1.4. Prevención en pérdida de datos

Es de rigor destacar que en la actualidad existe una gran conexión entre la investigación dentro de la disciplina del procesamiento del lenguaje natural (NLP) y la investigación en prevención de pérdida de datos. En los últimos años, los estudios en seguridad de redes informáticas empezaron a abordar el problema de la detección automática de ataques desconocidos en el instante que alcanzan al sistema objetivo. Estos ataques tratan de explotar la semántica de la comunicación de red entre el cliente y el servidor de aplicaciones, a fin de obtener acceso a través del ordenador atacado o por lo menos para evitar que se trabaje normalmente. El proceso de comunicación definidos por los protocolos de la capa de aplicación (por ejemplo HTTP, FTP, RPC o IMAP) también puede ser considerada como una comunicación basada en texto en una lengua artificial [37].

El análisis de la carga útil tratando los datos como secuencias de bytes ha sido estudiado por diversos autores. Wang y Stolfo [120] presentaron PAYL un detector automático de intrusiones, basado en aprendizaje no supervisado. En primer lugar, PAYL calcula durante una fase de entrenamiento un perfil de distribución de frecuencias de bytes y su desviación estándar de la carga que fluye de un solo host a un puerto. A continuación, durante la fase de detección PAYL calcula la similitud de los nuevos datos respecto del perfil pre-calculado utilizando la distancia Mahalanobis [73]. El detector compara esta medida con un umbral y genera una alerta cuando la distancia de una nueva entrada supera este umbral.

Más tarde, Wang et. al diseñaron Anagrama [119], que es un detector de anomalías de contenido que modela una mixtura de orden superior de n-gramas ($n > 1$) diseñado para detectar anomalías y "sospechosas" cargas útiles de paquetes de red.

Gracias a que las investigaciones en detección en seguridad en redes se están centrado en la aplicación de métodos de aprendizaje automático más avanzados, la generalización de la extracción y la representación de las características ha aumentado mucho la flexibilidad de la definición de medidas de similitud entre datos secuenciales, dentro de un contexto de seguridad. El trabajo de Rieck y Laskov [96] presenta una forma eficaz de combinar características extraídas de las secuencias de bytes, por ejemplo, palabras, n-gramas con un

[3] http://www.illc.uva.nl/EuroWordNet/

valor n arbitrario o secuencias contiguas, para una amplia gama de medidas de similitud lineales y no lineales.

2.2. Herramientas disponibles para un empresa

Actualmente, hay disponibles herramientas de detección automática de plagio que una empresa puede utilizar para protegerse. Todas estas herramientas utilizan métodos con referencias. A continuación, exponemos algunas de las herramientas disponibles más conocidas:

2.2.1. Turnitin

Turnitin[4] es una herramienta de pago para detectar automáticamente documentos sospechosos de cometer plagio que fue desarrollada por el Dr. John Barrie de la Universidad de Berkeley, California, y es utilizado por más de 50 universidades de todo el mundo [2].

Turnitin está dirigido hacia el área académica. Esta herramienta permite a los profesores verificar que los trabajos de los estudiantes tienen la citación adecuada o comprobar que no se comete plagio mediante la comparación con una bases de datos que se actualiza continuamente. Al finalizar el proceso Turnitin presenta un informe con los resultados el cual proporciona a los instructores la oportunidad de enseñar a sus alumnos los métodos adecuados de citación, así como salvaguardar la integridad académica de sus estudiantes [2].

En 2004, un grupo de universidades australianas decidieron realizar unos experimentos para explorar las percepciones de los profesores de la utilidad y aplicabilidad de Turnitin en las aulas terciarias. Para la realización de los experimentos eligieron a 2.000 estudiantes y siete profesores, los cuales fueron ubicados en los cuatro campus de la Universidad de la Costa Sur (SCU) y los siete profesores ejercían en los programas universitarios de cinco facultades diferentes: Artes, Negocios, Derecho, Educación, Ciencia y Tecnología. Este estudio puso de relieve las ventajas de usar software de detección de plagio, como Turnitin. Mientras que algunos consideraron que era beneficioso para la detección de coincidencia en documentos de texto, y que ahorraba tiempo. Otros temían que podría ser utilizado como medio de castigar a los alumnos dentro de la política y que el problema subyacente de plagio sigue sin resolverse. Finalmente, se consideró que Turnitin podría desempeñar una función muy útil para concienciar a los estudiantes en el problema del plagio como una cuestión de integridad académica [114].

[4] http://www.turnitin.com/

2.2.2. WCopyFind

WCopyFind[5] es un software desarrollado por Bloomfield de la Universidad de Virginia (2004). WCopyFind detecta plagio realizando una búsqueda a través de la comparación de n-gramas. El tamaño de los n-gramas es proporcionado por el usuario, aunque Bloomfield sugiere hexagramas como tamaño ideal. En la figura 2.1 muestra la interfaz de la herramienta WCopyFind.

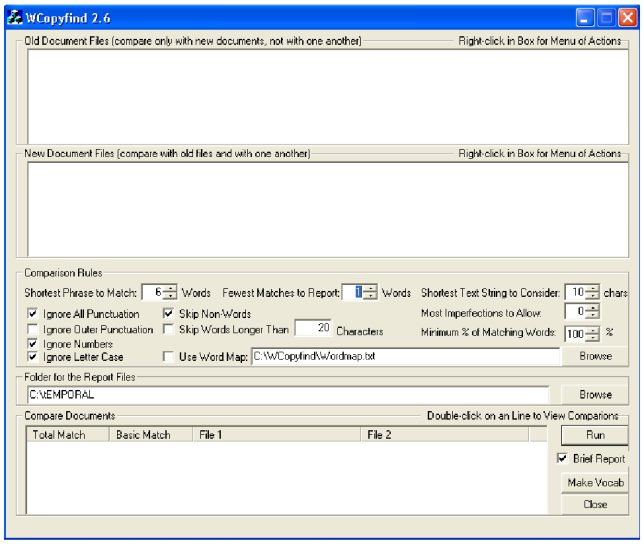

Figura 2.1: Interfaz de la herramienta WCopyFind

Debido a que WCopyFind utiliza como unidades de comparación n-gramas para detectar plagio, el idioma de los textos no es importante siempre y cuando los documentos comparados estén escritos en el mismo idioma. Hay que destacar que este sistema no puede encontrar el plagio en documentos no introducidos en él, ya que es un sistema de detección de plagio con referencia. Por lo que no se puede buscar plagio en textos de ninguna fuente externa, a menos que incluya esa fuente externa en los documentos que le asigne WCopyFind. Es decir, que funciona en datos puramente locales, no puede buscar en la Web o en Internet para encontrar los documentos correspondientes. Si el usuario tiene sospechas que una fuente externa, ha cometido plagio, debe crear un documento local que contenga dicho texto e incluir este documento en la colección de documentos de WCopyFind.

Para determinar la eficacia de WCopyfind el autor realizó unos experimentos, donde 600 trabajos de los estudiantes de un curso sobre el impacto social de la tecnología de la información fueron revisados en busca de posibles plagios. Los trabajos tenían entre 500 y 2000 palabras. El sistema demostró ser computacionalmente muy eficiente y tardó poco tiempo en marcar cinco casos que requerían un examen más detallado [22].

[5]http://plagiarism.phys.virginia.edu/

2.2.3. Ferret

Ferret[6] es una herramienta para detectar plagio desarrollada por la Universidad de Hertfordshire [70]. Éste analiza documentos proporcionados por el usuario de diferentes formatos, como son PDF, Word y RDF. Además ofrece una interfaz muy sencilla e intuitiva al usuario. Ferret detecta plagio en varias lenguas, tanto en lenguaje natural como lenguajes de programación. Además el algoritmo utilizado tiene unos costes, tanto temporal como computacional, lineales [74].

Ferret trabaja con la extracción de trigramas, obteniendo una medida de similitud que se basa en el porcentaje de trigramas comunes que existe entre un par de documentos [74], es decir, si A es el conjunto de trigramas que pertenece al documento 1, y B es el conjunto de trigramas que pertenece al documento 2, entonces:

$$Similitud = \frac{\text{Número de trigramas comunes}}{\text{Número total de trigramas}} = \frac{|A \cap B|}{|A \cup B|} \tag{2.8}$$

En definitiva, Ferret es una herramienta de escritorio para detectar plagio, lo que significa que tiene que ser rápido e interactivo. Tiene que ser rápido, porque un ser humano está a la espera de los resultados; y tiene que ser interactivo, porque el factor humano está disponible, además de ser apropiado [75], ya que en última instancia la decisión de si es o no plagio debe ser tomada por profesionales.

2.2.4. iThenticate

iThenticate[7] es un servicio de detección de plagio para el mercado corporativo, siendo una extensión de iParadigms. El servicio fue lanzado en 2004, como resultado de la demanda del mercado. La Organización Mundial de la Salud fue la primera en adoptar iThenticate, pero el grupo de usuarios ha crecido para incluir organizaciones en muchos sectores, tales como editoriales, medios de comunicación, centros de investigación, agencias gubernamentales, instituciones financieras y firmas legales.

iThenticate ofrece tres servicios principales: prevención del plagio basado en la Web y verificación de contenido; protección de la propiedad intelectual, comparando su contenido a su extensa base de datos, y prevención contra la apropiación indebida; y comparación de documento-a-documento para la búsqueda de concordancia de textos.

El servicio presenta un informe de similitud que muestra las coincidencias del documento presentado con los documentos de la base de datos de iThenticate. Los informes de similitud incluyen:

- Comparaciones directas de palabras coincidentes de documentos relacionados;

[6] http://homepages.feis.herts.ac.uk/∼pdgroup/
[7] http://www.ithenticate.com/

- Reconocimiento de patrones coincidentes de ambos documentos;
- Capacidad para ver todas las palabras plagiadas subyacentes que han sido oscurecidas por la superposición de plagios;
- Opción de volver a enviar el informe de similitud para incluir todos los datos actuales en la base de datos;
- Opción de crear una base de datos de documentos de específicas organizaciones para generar una base de datos interna para informes de similitud.

2.2.5. Plagiarism Checker

Plagiarism Checker[8] es una aplicación web creada por el Departamento de Educación de la Universidad de Maryland. Esta aplicación detecta textos sospechosos de ser plagio. El funcionamiento es sencillo, hay que copiar el texto del cual sospechamos, introducirlo en la caja de texto, y comprobar si encuentra algo similar o igual por la red. En la figura 2.2 su puede ver la interfaz de Plagiarism Checker.

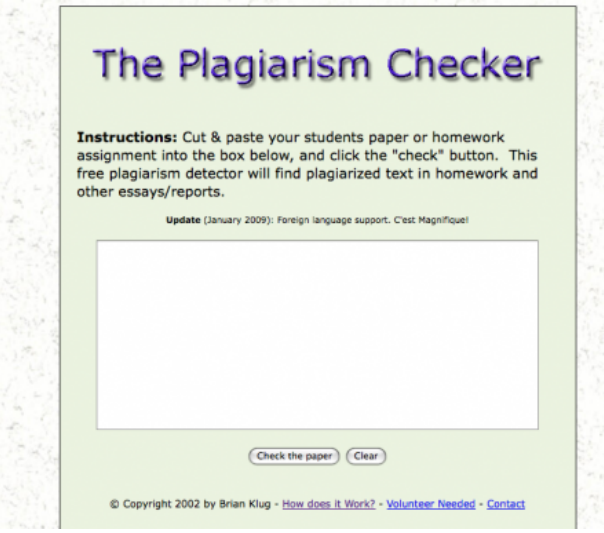

Figura 2.2: Interfaz de Plagiarism Checker

El servicio es totalmente gratuito, no requiere registro, y es bastante rápido. Plagiarism Checker utiliza la API de Google[9] para buscar los posibles textos copiados, con lo que tenemos la seguridad de que si existe algo en la red, lo encontrará.

[8] http://www.dustball.com/cs/plagiarism.checker/
[9] http://www.google.com/

2.2.6. DOC Cop

DOC Cop[10] es una herramienta libre de licencia y gratuita. Se trata de un proceso que proporciona resultados aceptables. Es mucho mejor en la comprobación de plagio entre los documentos presentados que para la comprobación de la presencia de material plagiado procedente de Internet. Si bien es gratuita, existen mejores soluciones disponibles [99].

Escanea documentos de MS Word unos contra otros en busca de similitudes. Cuando son menos de 10 documentos el límite es de 250.000 palabras por cada documento, en caso contrario el límite es de 12.000 palabras. La respuesta se envía por correo electrónico destacando los fragmentos de texto donde se sospecha que pueden contener plagio.

2.2.7. Pl@giarism

Pl@giarism[11] es una herramienta gratuita y semi-automática desarrollada por la Universidad de Maastricht (Bélgica). La Facultad de Leyes de la Universidad de Maastricht utilizan esta herramienta para la detección de plagio en los ensayos de los estudiantes.

Pl@giarism es un simple programa para Windows que automatiza el proceso de determinar la similitud entre pares de documentos comparando trigramas. La herramienta busca en documentos locales o en un servidor con formato MS-Word, docuemtos web residentes. Pl@giarism devuelve un base de datos en MS-Access con tablas indicando varias emparejamientos entre documentos con sus respectivos porcentajes de similitud. En la figura 2.3 se puede ver el informe de resultados que muestra la herramienta Pl@agiarism. Como se observa se puede seleccionar todos las similitudes con todos los documentos o seleccionar los documentos a comparar.

2.2.8. CopyCatch

CopyCatch[12] es una herramienta para detección automática de plagio diseñada por CFL Software. Este herramienta está diseñada para encontrar similitudes entre documentos utilizando documentos completos, oraciones o frases. CopyCatch necesita la introducción de los documentos electrónicos para realizar la búsqueda de posible plagio. Los algoritmos se basan principalmente en el uso de herramientas del lenguaje natural más que en modelos estadísticos o matemáticos. Además, es posible utilizarlo en varios idiomas. Sus algoritmos se basan en comparaciones que se encuentran por debajo del nivel de similitud de sentencia y están diseñados para ser capaces de identificar la similitud incluso cuando hay cambios en el orden de las palabras, inserciones o eliminaciones.

CopyCatch tiene diversas versiones:

[10] http://www.doccop.com/
[11] http://www.plagiarism.tk/
[12] http://cflsoftware.com/

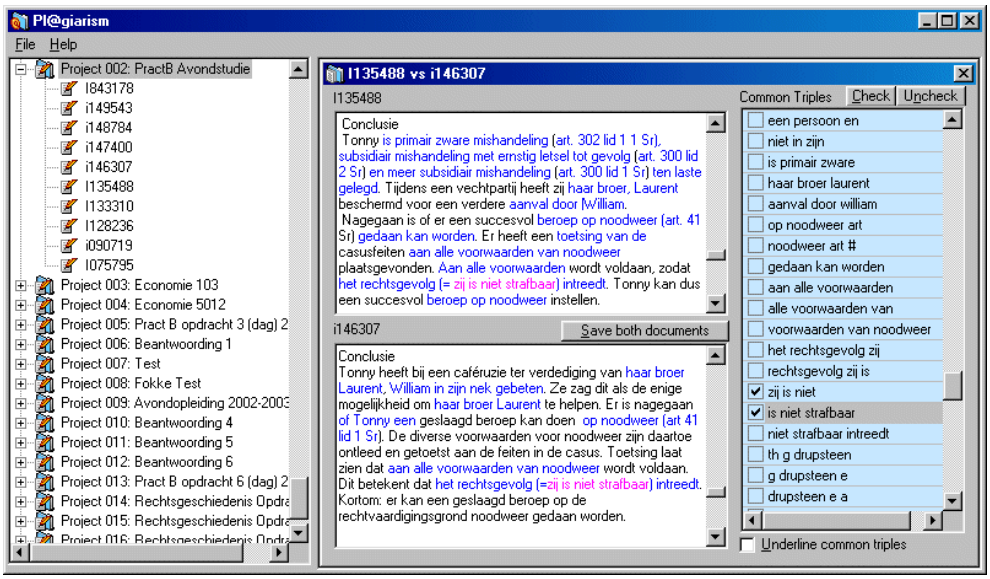

Figura 2.3: Informe de resultados de Pl@giarism

- CopyCatch Gold: utilizado por profesores universitarios para supervisar el trabajo del estudiante.

- CopyCatch Investigator: es un conjunto de algoritmos diseñados para usuarios a gran escala; normalmente se adaptan con fines específicos, desde las interfaces visuales hasta los sistemas de control automatizados que se ejecutan en potentes ordenadores con multiprocesadores.

- CopyCatch Analyst: es un conjunto de programas que en conjunto ofrecen detallados análisis numéricos, estadísticos, frases y vocabulario de los textos.

- CopyCatch Campus: fue desarrollado en cooperación con la Open University, que había decidido utilizar CopyCatch Gold, pero necesitaba ampliar sus capacidades para manejar cursos de miles de personas y automatizar todo el proceso.

- CopyCatch Citereader: lee los documentos y puede reconocer citas debidamente referenciadas en éstos. En caso contrario, el usuario obtiene una indicación de las frases que valdría la pena buscar la fuente.

- CopyCatch Summariser: es un sistema automático que realiza resúmenes. Utiliza un enfoque diseñado para producir un resumen de una extensión y contenido específico, que puede ser variado por el usuario.

2.2.9. EVE2

EVE2[13] (Essay Verification Engine) es un herramienta desarrollada por Canexus. EVE2 es una herramienta muy potente que permite a los profesores y maestros en todos los niveles del sistema educativo para determinar si los estudiantes han plagiado material de la World Wide Web. EVE2 acepta ensayos en texto plano, en formato Microsoft Word o Corel Word Perfect y devuelve enlaces a páginas desde las que puede un estudiante haber plagiado. EVE2 ha sido desarrollado para ser lo suficientemente potente como para encontrar material plagiado llevando a cabo un gran número de búsquedas en Internet.

Sin embargo, Dreher en [22] realizó un experimento para localizar posibles textos plagiados en 16 páginas que contenían unas 7.300 palabras de una tesis de master. EVE2 tardó alrededor de 20 minutos para completar la tarea. Finalmente, Dreher argumenta que esta herramienta era tan lenta que sólo se podría verificar algunos artículos cuidadosamente seleccionados [22].

2.2.10. MyDropBox

MyDropBox[14] es un servicio "online" que facilita la detección del plagio para administradores, profesores y estudiantes. Los informes generados están bien estructurados y codificados por color para facilitar la vinculación de texto de nuevo a las fuentes de Internet [99]. Entre otras funcionalidades tiene:

- Producción de los informes de la originalidad en relación con:
  - Fuentes de Internet: *MSN Search Index* proporcionado por Microsoft Corporation y contiene más de 8 millones de documentos actualizados continuamente;
  - Base de datos de los documentos presentados anteriormente;
- Posee un asistente que ayuda en la correcta producción de referencias

[13] http://www.canexus.com/
[14] http://www.mydropbox.com/

Capítulo 3

Evaluación en la competición PAN

Así como comentamos en la introducción, la Web 2.0 ha suspuesto una revolución en el mundo empresarial mejorando las relaciones entre empresas y clientes. Sin embargo, la facilidad de acceso a la información que proporciona la Web ha aumentado la competencia desleal entre empresas, plagiando sus ideas, productos o servicios. Además, si sumamos que gracias a la gran cantidad de información que circula por la Web, las empresas desleales se sienten seguras ante la dificultad de ser descubiertas, comprenderemos como es posible que algunas empresas utilicen el plagio como práctica habitual.

Cuando a una empresa le copian sus ideas, sus productos, sus principios, sus servicios o sus novedades, pierde el distintivo que la diferencia del resto de empresas. Es decir, aquello por lo que ha estado trabajando durante cierto tiempo, puede que incluso años, se pierda porque no ha tenido la previsión de proteger su propiedad intelectual. Por este motivo las empresas están obligadas a protegerse.

Así como describimos en el capítulo anterior, actualmente, existen herramientas disponibles que pueden utilizar las empresas para detectar automáticamente plagio. Una de estas herramientas es WCopyFind, que ya hemos mencionado anteriormente en el Capítulo 2 Sección 2.2. Para comprobar la eficacia de esta herramienta hemos participado en la competición *1st International Competition on Plagiarism Detection* (PAN'09).

En este capítulo hablaremos de experimentos realizados en la participación de la competición PAN'09. El capítulo esta organizado como sigue: en la Sección 3.1 describimos el corpus utilizado en los experimentos, en la Sección 3.2 describimos cual es la tarea que debíamos realizar en la competición, en la Sección 3.3 comentamos las diferentes medidas para verificar la eficacia de las herramientas y en la Sección 3.4 discutimos los resultados obtenidos en los experimentos.

3.1. Descripción del corpus

Para los experimentos, hemos utilizado el corpus que proporcionaban en la competición de PAN'09. Este corpus estaba formado principalmente por documentos en inglés, en los cuales se pueden encontrar cualquier tipo de plagio, desde copias exactas de fragmentos hasta fragmentos traducidos desde el alemán o el castellano. Este corpus fue generado automáticamente por los responsables de la competición utilizando una aplicación que introducía fragmentos plagiados en documentos de una forma aleatoria.

Esta aplicación escogía un texto y decidía si iba a introducir algún fragmento plagiado o no, de qué documentos plagiar, cuántos pasajes se plagiaban, y qué tipo de plagio y longitud se trataba. También decidía si el fragmento plagiado sería traducido desde otro idioma. Los fragmentos plagiados se realizaban mediante una secuencia aleatoria de operaciones de textos en las que se incluye la eliminación de palabras, la inserción de palabras de una fuente externa, copia exacta, o la sustitución de palabras por sinónimos, antónimos, hiperónimos o hipónimos. Los fragmentos plagiados traducidos se creaban a partir de traducción automática.

Este corpus se encontraba en formato plano codificado con UTF-8. Los documentos se encontraban divididos en dos carpetas diferentes, en una se encontraban los documentos sospechosos de ser plagios y en la otra se encontraban los documentos originales.

En total disponíamos de 7.214 documentos sospechosos, los cuales podían ser plagios de uno o más documentos originales, o no contener ningún fragmento plagiado. Por otro lado, existían otros 7.215 documentos originales. El tamaño de los documentos, tanto en los sospechosos como en los originales, variaban entre textos cortos y documentos largos. De esta forma se podía comprobar la eficacia de un detector de plagio, ya que no se restringe la búsqueda en textos cortos o largos.

3.2. Descripción de la tarea de detección con referencia

La competición PAN'09 estaba formada por dos tareas diferentes: una consistía en la detección de plagio con referencia y la otra consistía en el análisis intrínseco del plagio. Como nuestro objetivo era investigar la eficacia de las herramientas disponibles que una empresa tendría a su alcance para poder detectar casos de plagio, como es la herramienta WCopyFind. Como la herramienta WCopyFind sólo es capaz de detectar plagio con referencias, hemos participado en la primera tarea de la competición.

La tarea consistía en dado un conjunto de documentos sospechosos y un conjunto de documentos originales, encontrar todos los pasajes de texto en los documentos sospechosos que han sido plagiados y los pasajes de texto correspondiente en los documentos originales.

Una vez detectados los fragmentos plagiados de un texto, se debía de incluir en un fichero XML todos los fragmentos plagiados, especificando los siguientes datos: la posición del carácter en el documento sospechoso donde comienza el fragmento, la longitud en caracteres

del fragmento en el documento sospechoso, el nombre del fichero original, la posición del carácter en el documento original donde comienza el fragmento y la longitud en caracteres del fragmento en el documento original. A continuación mostramos el formato del fichero XML:

```
<document reference="...">                <!-- 'reference' refers to the analysed suspicious document -->
  <feature name="detected-plagiarism"     <!-- plagiarism which was detected in an external analysis -->
           this_offset="5"                <!-- the char offset within the suspicious document -->
           this_length="1000"             <!-- the number of chars beginning at the offset -->
           source_reference="..."         <!-- reference to the source document -->
           source_offset="100"            <!-- the char offset within the source document -->
           source_length="1000"           <!-- the number of chars beginning at the offset -->
  />
  ...                                     <!-- more external analysis results in this suspicious document -->

  <feature name="detected-plagiarism"     <!-- plagiarism which was detected in an intrinsic analysis -->
           this_offset="5"                <!-- just like above but excluding the "source"-attributes -->
           this_length="1000"
  />
  ...                                     <!-- more intrinsic analysis results in this suspicious document -->
</document>
```

3.3. Medidas de evaluación

Para la medición del éxito de la herramienta de detección de plagio teníamos en cuenta la precisión, el recall y la granularidad en la detección de los pasajes plagiados en el corpus. En todas las ecuaciones para el cálculo del éxito de la herramienta, hay que tener en cuenta la siguiente notación: s denota un fragmento plagiado del conjunto S de todos los fragmentos plagiados; r que denota la detección del conjunto R de todas las detecciones; S_R indica el subconjunto de S para los que existen detecciones en R; $|s|$, $|r|$ representan la longitud en caracteres de s, r y $|S|$, $|R|$, $|S_R|$ denotan los tamaños de los conjuntos respectivos. Las fórmulas para el cálculo del éxito son de la siguiente manera:

$$recall = \frac{1}{|S|} \sum_{i=1}^{|S|} \frac{\alpha(s_i)}{|s_i|} \tag{3.1}$$

donde $\alpha(s_i)$ representa el número de caracteres detectados de s_i.

$$precision = \frac{1}{|R|} \sum_{i=1}^{|R|} \frac{\beta(r_i)}{|r_i|} \tag{3.2}$$

donde $\beta(r_i)$ representa el número de caracteres plagiados de r_i.

$$granularidad = \frac{1}{|S_R|} \sum_{i=1}^{|S_R|} \gamma(s_i) \tag{3.3}$$

donde $\gamma(s_i)$ representa el número de detecciones de s_i en R.

$$total = \frac{F}{\log_2(1 + granularidad)} \tag{3.4}$$

donde F es la medida harmónica (*F-measure*) de recall y precisión, es decir:

$$F = \frac{w * precision * recall}{precision + recall} \tag{3.5}$$

3.4. Discusión de los resultados

La herramienta WCopyFind realiza la comparación a partir de n-gramas. Sin embargo, la herramienta permite seleccionar el tamaño de los n-gramas. Por tanto, antes de participar en la competición hemos realizado varios experimentos con el corpus de entrenamiento para decidir qué tamaño era el más indicado. Debemos señalar que el corpus de entrenamiento tenía las mismas características que el corpus de la competición.

En estas pruebas hemos decidido experimentar con tetragramas, pentagramas y hexagramas. Para decidir qué tamaño era el más indicado para la tares, hemos tenido en cuenta la medida de *F-measure*. En la tabla 3.1 mostramos los resultados obtenidos en cada una de las pruebas:

n-grama	Precisión	Recall	F-measure
4	0,0020	0,4103	0,0041
5	0,0801	0,3601	0,1310
6	0,0924	0,3423	0,1455

Tabla 3.1: Resultados en la fase de entrenamiento

En la tabla 3.1 podemos destacar varios puntos interesantes. Por un lado hay que destacar que, contrariamente a otras tareas de ingeniería del lenguaje, la medida de precisión es inferior a la medida de recall. Esto se debe principalmente al tamaño del corpus, que está formado por una gran cantidad de documentos, de modo que existe una probabilidad muy alta de que se encuentre un fragmento de un documento sospechoso similar entre los documentos originales, sin ser un fragmento plagiado.

Estas pruebas nos han confirmado un rasgo característico de la tarea de la detección del plagio con referencia. Analizando la tabla 3.1 se observa que cuánto más pequeño el tamaño de n-gramas, menor es la precisión. Esto se debe porque cuando los n-gramas son muy pequeños es más probable que encuentre algún fragmento similar en cualquier otro documento.

Software	Precision	Recall	F-measure	Granularidad	Total
Encoplot	0,7418	0,6585	0,6976	1,0038	0,6957
WCopyFind	0,0136	0,4586	0,0265	1,0068	0,0264
Ferret	0,0290	0,6048	0,0553	6,7780	0,0187

Tabla 3.2: Resultados obtenidos en la competición PAN'09

Sin embargo, sucede todo lo contrario con la medida de recall, es decir, cuánto más pequeño son los n-gramas es mayor el recall. Este hecho puede explicarse por el mismo motivo que con los resultados en la medida de precisión, es decir, cuánto más pequeños son los n-gramas, hay una mayor probabilidad de encontrar fragmentos similares en el documento plagiado, y será mucho más probable encontrar los fragmentos plagiados en realidad.

Después de completar las pruebas en la fase de entrenamiento, hemos tomado la decisión de que el mejor tamaño para los n-gramas es de hexagramas, ya que es el que mejor resultado hemos obtenido en la medida de *F-measure*. Además, coincide con el tamaño que asesora al desarrollador de la aplicación de la herramienta WCopyFind. Sin embargo, difiere con la conclusión a la que llegaron Lyon y sus colaboradores en [69] o Barrón y Rosso en [6], ya que ambos estudios sugieren que trigramas son la medida idónea.

Una vez decidido el tamaño idóneo de los n-gramas, hemos realizado los experimentos para la competición. La tabla 3.2 muestra los resultados que hemos obtenido con el corpus de la competición con la herramienta WCopyFind [118]. También muestra los resultados obtenidos por el grupo formado por Malcolm, Lane y Rainer [75], que utilizó la herramienta de Ferret [93].

Observando los resultados, podemos comprobar que para ambas herramientas, los resultados no son buenos. Queremos hacer hincapié en que los resultados de la medida de precisión son muy bajos. Como ya hemos discutido, esto se debe principalmente al tamaño del espacio de búsqueda.

Otra de las causas que afectan al bajo rendimiento en la competición de WCopyFind, es que esta herramienta no puede encontrar plagio cuando hay traducciones a idiomas diferentes al del documento original. Otro factor desfavorable añadido que tiene WCopyFind es que tampoco tiene en cuenta la modificación de palabras, como pueden ser sinónimos, antónimos, hiperónimos o hipónimos.

Otro aspecto a destacar es el tiempo de ejecución de WCopyFind. En la competición, hemos comparado cada documento sospechoso con todos los documentos originales, por tanto el número total de comparaciones que ha realizado la herramienta ha sido de: 7.214 x 7.215 = 52.049.010. Además, el tamaño de los ficheros variaba entre 1 KB y 2,5 MB, es decir, existían ficheros muy extensos provocando que la comparación fuera muy lenta. La duración total de todo el análisis del corpus ha sido de tres semanas repartiendo el trabajo en dos computadoras diferentes. Es decir, un tiempo excesivo.

Sin embargo, una empresa no puede permitirse el lujo de tener dos computadoras exclusivamente para detectar plagio. Además, debemos indicar que aunque 52 millones de comparaciones parece un número muy alto, en realidad no son nada en comparación con la cantidad de documentos que discurren por la Web.

Por tanto las pruebas realizadas con las herramientas disponibles para la detección de plagio, como WCopyFind y Ferret, indican tanto por los resultados obtenidos, así como por el tiempo de ejecución, que no son una buena opción para que las empresas los utilicen para comprobar que sus documentos no han sido plagiados por otras empresas, puesto que la cantidad de documentos que existen hoy por hoy en la Web es tan enorme que con la precisión de estas herramientas devolverían tal cantidad de documentos sospechosos de cometer plagio, que una persona no podría analizarlos para verificar el plagio.

Esto nos muestra la dificultad que tiene hoy en día las empresas para proteger su propiedad intelectual. Es por esto que es necesario el desarrollo de métodos ad-hoc de detección autmoática de plagio para que las empresas puedan proteger su propiedad intelectual.

Capítulo 4

Análisis de opiniones con ontologías

El análisis de opiniones (o minería de opiniones) es una técnica de procesamiento de lenguaje natural que identifica la orientación semántica de una opinión. Una opinión sobre un objeto puede tener diferentes polaridades: positiva, negativa y neutra. Por ejemplo, en una opinión de un hotel puede describirse como *maravilloso*, mientras que en otra opinión el mismo hotel puede describirse como *desastroso*; el primer término tiene una connotación positiva y el segundo, en cambio una connotación negativa.

Es decir que, dado un conjunto de documentos de textos evaluativos D que contienen opiniones (o sentimientos) acerca de un objeto, la minería de opiniones tiene como objetivo extraer las características y los componentes del objeto que han sido objeto de comentarios en cada documento $d \in D$ y determinar si los comentarios son positivos, negativos o neutrales [67].

En la actualidad la Web 2.0 ha propiciado un nuevo concepto de trabajo centrado en las opiniones de los consumidores. Una de las nuevas tendencias de las empresas es investigar métodos automáticos capaces de detectar opiniones plagiadas en los blogs. Esto es debido a que la relevancia de un blog se mide a partir de las visitas y de la cantidad de opiniones que se generan en ellas, por tanto, al aumentar las opiniones generadas en una blog, aumenta su relevancia (o influencia), y esto puede traducirse con beneficios publicitarios.

Sin embargo, no existe en la actualidad un corpus para el análisis del plagio de opiniones, debido a eso hemos decidido investigar la integración (legal) de análisis de opiniones vía fusión de ontologías. Otro campo de interés para las empresas en la actualidad es la integración de análisis de opiniones. Esto se debe a que analizar las opiniones de los consumidores generadas en los blogs supone obtener información de primera mano sobre las virtudes y defectos de los productos de la empresa, así como las tendencias del mercado. No obstante, analizar toda la información de la Web 2.0 es una tarea irrealizable, es por ello que actualmente las empresas tienen mucho interés en poder compartir e intercambiar entre ellas la información del análisis de opiniones.

El capítulo está organizado de la siguiente manera: en la sección 4.1 describimos el estado

del arte del análisis de opiniones, en la sección 4.2 describimos los trabajos enfocados al análisis de opiniones basados en ontologías y en la sección 4.3 exponemos el algoritmo propuesto para el análisis de opiniones vía fusión de ontologías, además en esta sección describimos el estado del arte de la disciplina de fusión de ontologías, ya que el algoritmo propuesto incluimos una fase de fusión de ontologías.

4.1. Estado del arte

La obra clásica por antonomasia dedicada a la medición del significado emotivo o afectivo en los textos es la Teoría de Diferenciación Semántica de Charles Osgood [88]. Osgood y sus colaboradores definen el significado emotivo como:

> *El significado emotivo es un aspecto estrictamente psicológico: los estados cognitivos del lenguaje humano de los usuarios que son condiciones necesarias a priori para la codificación selectiva de los signos léxicos y condición necesaria a posteriori en la descodificación selectiva de signos en los mensajes*[1].

La Teoría de Diferenciación Semántica consiste en utilizar varios pares de adjetivos bipolares para ampliar las respuestas de sujetos a palabras, frases cortas, o textos. Es decir, a diferentes personas se les pidió que calificaran el significado de éstos en diferentes escalas; tales como: activo/pasivo, bueno/malo, optimista/pesimista, positivo/negativo, fuerte/débil, serio/gracioso, y feo/bonito.

Cada par de adjetivos bipolares es un factor en la Teoría de Diferenciación Semántica. Como resultado, la Teoría de Diferenciación Semántica puede hacer frente a todo un gran número de aspectos del significado afectivo. Es lógico preguntarse si cada uno de estos factores es igualmente importante. Osgood *et al.* [88] utilizaron el análisis factorial en extensas pruebas empíricas para investigar dicha cuestión, obteniendo una sorprendente respuesta puesto que la mayoría de la variación de los datos podría explicarse por tres factores principales. Estos tres factores del significado afectivo o emocional son el factor de evaluación (por ejemplo, bueno/malo, bonito/feo, bueno/cruel y honesto/deshonesto), el factor de potencia (por ejemplo, fuerte/débil, grande/pequeño y pesado/ligero), y el factor de actividad (por ejemplo, activo/pasivo, rápido/lento y frío/calor). Entre estos tres factores, el factor de evaluación tiene la mayor importancia relativa. El análisis de opiniones está orientado al factor de evaluación [116].

Los trabajos realizados hasta la actualidad sobre análisis de opiniones se pueden dividir básicamente en dos enfoques: estudios basados en diccionarios y basados en corpus. En los apartados siguientes explicamos cada uno de estos enfoques.

[1] Para más información ver [88], página 318

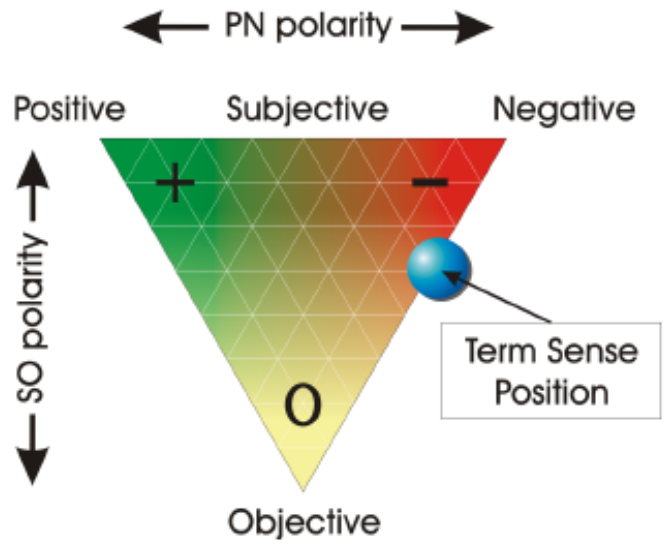

Figura 4.1: Representación gráfica de SentiWordNet

4.1.1. Basados en diccionarios

Los métodos basados en diccionarios utilizan tesauros para clasificar la orientación semántica de textos evaluativos. Uno de los recursos más representativos de este grupo de estudios es SentiWordNet[2] [24] el cual es un recurso léxico en el cual cada *synset* de WordNet[3] está asociado con tres puntuaciones numéricas: $Obj(s)$, $Pos(s)$ y $Neg(s)$, correspondientes al valor como palabra objectiva, positiva y negativa. En la figura 4.1 se puede ver una representación gráfica de SentiWordNet. Otro recurso representativo es WordNet-Affect[4] [113] el cual es una jerarquía adicional de etiquetas de dominio afectivo, las cuales representan los conceptos afectivos más comentados.

Sin embargo, ni SentiWordNet ni WordNet-Affect son los primeros recursos léxicos generados. En la década de 1960, Stone y Lasswell comenzaron la construcción de léxicos en el cual las palabras estaban etiquetadas con afectos. En *Laswell Value Dictionary* [60], la palabra *admire*, por ejemplo, fue etiquetada con un valor positivo dentro de la dimensión de *respect*. En este diccionario estaban etiquetadas las palabras con valores binarios dentro de ocho dimensiones básicas: *wealth* (riqueza), *power* (poder), *rectitude* (rectitud), *respect* (respeto), *enlightenment* (iluminación), *skill* (habilidad), *affection* (afecto) y *wellbeing* (bienestar). El trabajo de Stone en el *Inquirer General Dictinoary* [111] ha continuado hasta nuestros días[5]. Actualmente, el diccionario contiene 1.915 palabras marcadas como generalmente positivas y 2.291 palabras como negativas. Una amplia variedad de otras clases de afectos se utilizan

[2]http://sentiwordnet.isti.cnr.it/
[3]http://wordnet.princeton.edu
[4]http://wndomains.itc.it/download.html
[5]En http://www.wjh.harvard.edu/∼inquirer/inqdict.txt se puede encontrar una versión de este diccionario

para etiquetar las entradas, por ejemplo, *pleasure* (placer), *pain* (dolor), *feel* (sentimiento), *arousal* (excitación), *emotion* (emoción), *virtue* (virtud) y *vice* (vicio)[6]. Por ejemplo, la palabra *admire* posee entre sus etiquetas *positive* y *pleasure*. En ambos diccionarios, todas las etiquetas son binarias, es decir, las palabras o bien poseen una cualidad, o no.

Wiebe [121] utilizó un conjunto de adjetivos "subjetivos" como semilla y un método de generación de tesauros [42] para encontrar más adjetivos "subjetivos". Turney y Littman [116] han demostrado que es posible descubrir automáticamente palabras cargadas positiva y negativamente, habida cuenta de catorce semillas de palabras, y utilizando las estadísticas de la asociación de la WWW. El conjunto de palabras positivas eran las semillas: {*good, nice, excellent, positive, fortunate, correct, superior*} (bueno, bonito, excelente, positivo, afortunado, correcto, superior) y su conjunto de palabras con carga negativa fueron las semillas: {*bad, nasty, poor, negative, unfortunate, wrong, inferior*} (malo, feo, pobre, negativo, desafortunado, mal, inferior). Encontraron que las palabras positivas tienden a asociar más a menudo con las palabras positivas que con negativas, utilizando un formulario de información mutua [12] y las estadísticas de páginas en las que aparecen las palabras en Altavista[7]. Por ejemplo, si uno desea decidir si una palabra como *fortuitous* es positiva o negativa, entonces se podría enviar solicitudes como "*fortuitous NEAR good*" y "*fortuitous NEAR bad*", mediante el operador *NEAR* y la facilidad de búsqueda avanzada de Altavista. Usando este método, se logró 98,2 % de exactitud con los 334 adjetivos más frecuentes en el conjunto de test de [39]. Finalmente, también llegaron a la conclusión que debería ser posible ampliar de forma fiable una lista de palabras marcadas como positivas y negativas, como las que se encuentran en el lexicón General Inquirer [111].

Además de los estudios de clasificación de la polaridad a nivel de palabras, existen estudios enfocados en realizar una clasificación a nivel de oración para determinar si la frase es subjetiva u objetiva y/o determinar si la oración pertenece a una opinión positiva o negativa [54, 122, 123]. Hatzivassiloglou y Wiebe en [40] estudian la clasificación a nivel de frase para determinar si una frase es objetiva o subjetiva, es decir, si expresa un hecho o una opinión positiva o negativa. Wiebe y Riloff en [122] distinguen frases subjetivas de objetivas. Kim y Hovy en [54] proponen un clasificador de orientación semántica de palabras y frases en inglés utilizando tesauros.

Por otro lado, existen estudios que utilizan glosas como WordNet [28] para buscar sinónimos y antónimos, y de este modo determinar la orientación semántica de las palabras en base a un conjunto de semillas de palabras. En [46, 54] se proponen dos métodos en los que utilizan un pequeño conjunto de semillas de palabras para encontrar sus sinónimos y antónimos en WordNet, y así predecir la orientación semántica de los adjetivos. En WordNet, los adjetivos se organizan en grupos bipolares como se muestra en la figura 4.2, donde comparten la misma orientación de sus sinónimos y a su vez, tienen la orientación opuesta de sus antónimos. Para asignar la orientación de un adjetivo, se busca el conjunto de sinónimos del adjetivo dado, es decir, el *synset* al que pertenece, y el conjunto de antónimos del adjetivo. Si conoce-

[6]Para más información véase http://www.wjh.harvard.edu/∼inquirer/homecat.htm
[7]http://www.altavista.com

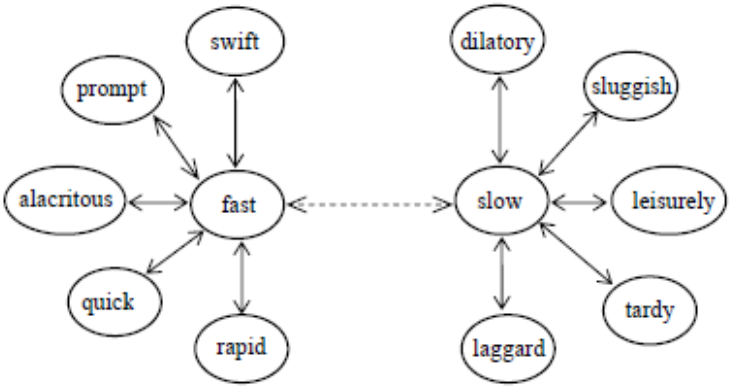

Figura 4.2: Estructura de adjetivos en grupos bipolares en WordNet

mos la orientación de un sinónimo o un antónimo, entonces la orientación del adjetivo dado podemos establecerlo a partir de éste. Como el grupo de sinónimos de un adjetivo siempre contiene un sentido que enlaza con la cabeza del grupo de sinónimos, el rango de búsqueda es bastante grande. Colocando un grupo de semillas de adjetivos con orientaciones conocidas suficientemente grande, se puede predecir la orientación semántica de todos los adjetivos [62].

4.1.2. Basados en corpus

El objetivo de los métodos basados en corpus es encontrar patrones de co-ocurrencia de palabras para determinar la orientación semántica de las palabras o frases [39, 115]. En este sentido se han propuesto varias técnicas para determinar la polaridad de palabras: desde medidas probabilísticas de asociación de palabras [116], así como técnicas que explotan la información sobre relaciones léxicas [51].

A finales de los 90 empezaron a realizarse trabajos que detectaban automáticamente la polaridad semántica de la información. Hatzivassiloglou y McKeown en [39] demostraron que, dado un conjunto de adjetivos emotivos, los adjetivos con una orientación positiva tienden a ser unidos con adjetivos con una orientación positiva, y los adjetivos negativos con los negativos, en expresiones como "bueno y honesto" o "malo y engañoso". En sus experimentos, decidieron que una serie de adjetivos frecuentes tenían algún tipo de orientación; posteriormente, utilizaron las probabilidades de que dos adjetivos apareciesen juntos en un corpus con el patrón "*X y Y*" para decidir si tenían la misma orientación. Ellos crearon grafos en los cuales los nodos eran los adjetivos y los enlaces entre nodos mostraban que los adjetivos aparecían con el patrón; posteriormente, dividieron este grafo en dos grupos con el mínimo número de enlaces entre los grupos. La clase más grande fue considerada como la clase formada por los adjetivos con polaridad negativa (ya que en inglés existen más palabras negativas que palabras positivas). En los experimentos obtuvieron un 92 % de precisión sobre un conjunto de test de 236 adjetivos que se clasificaron como positivos o negativos.

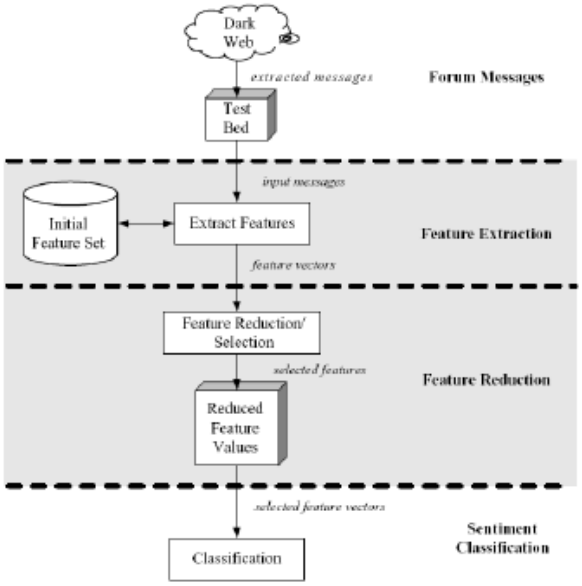

Figura 4.3: Diseño del sistema de análisis de sentimientos propuesto por Abbasi et al.

Abbasi et al. [3] propusieron el uso de metodologías de análisis de sentimientos para la clasificación de las opiniones realizadas en los foros de la Web 2.0 en varios idiomas (ver figura 4.3). El diseño tenía dos fases principales: la extracción de un conjunto inicial de características y, a continuación, la selección de características. Los experimentos muestran grandes resultados sobre un corpus de opiniones de películas. Este método se centra únicamente en la clasificación a nivel de documento.

Gamon et al. [31] presentaron un prototipo de sistema llamado *Pulse*, para tareas de minería y de orientación semántica de comentarios de clientes. Sin embargo, esta técnica está limitada al dominio de los productos y depende en gran medida del conjunto de entrenamiento, así que no es de aplicación general para clasificación de opiniones sobre un tema arbitrario.

Yu et al. en [125] propusieron un método híbrido para combinar *HowNet*[8] [21] y clasificadores de orientación semántica. Ellos dividieron las opiniones, en palabras y frases características extraídas de los datos de entrenamiento. Posteriormente, calculan la similitud semántica de las palabras y frases características con las palabras marcadas en *HowNet*, y adoptan los términos positivos o negativos como características del clasificador de orientación semántica. Además, se agregan reglas al clasificador para las frases negadas. Si se encuentra una palabra etiquetada como negación, se cambia la orientación de toda la frase. Sin embargo, el rendimiento del método no es satisfactorio de acuerdo con los resultados obtenidos en sus experimentos.

[8] http://www.keenage.com/

4.2. Análisis de opiniones basado en ontologías

Sin embargo, en los estudios enfocados a la clasificación de los textos de opinión tanto a nivel de documento o como a nivel de frase, no identifican cual es el objeto de la opinión favorable o desfavorable. Un documento positivo en un objeto no significa que el titular de la opinión tiene opiniones positivas sobre todos los aspectos o características del objeto. Del mismo modo, un documento negativo no significa que el titular de la opinión no le gusta todo lo relacionado con el objeto. En un documento de opinión, tales como una opinión del cliente de un producto, el titular de la opinión escribe tanto aspectos positivos como negativos del objeto, aunque el sentimiento general del objeto puede ser positivo o negativo. Se han propuesto varios estudios que extraen la opinión de cada concepto del que se opina para resumir la opinión general [46, 67, 91]. En la minería de opiniones basado en características, las características en términos generales significa las características del producto o atributos y funciones. Las principales tareas en esta técnica son:

- Identificar las características que han sido comentadas;
- Identificar si los comentarios son positivos o negativos.

Hu y Liu [46] propusieron un método para resumir la opinión de productos clasificados según la polaridad de la opinión del mismo. Hu y Liu propusieron la minería de asociación de palabras para extraer las características. A continuación, extraían los adjetivos de las frases que contienen al menos una característica extraída anteriormente. Por último, se genera un resumen con los pares característica-adjetivo de acuerdo con las características extraídas (véase figura 4.4). Ellos identificaron la orientación semántica de los adjetivos por medio de los sinónimos y antónimos obtenidos a través de WordNet.

Posteriormente, Liu *et al.* [67] mostraron un método para realizar resúmenes de las opiniones del estilo gráfico de barras, clasificados por las características del producto. Este modelo da una formulación más completa del problema de la minería de opiniones. En él se identifican las piezas clave de información que deben ser extraídas, y describe cómo un resumen estructurado de la opinión puede ser producido a partir de textos no estructurados. Sin embargo, tanto el trabajo de Hu y Liu [46] como el de Liu *et al.* [67] son dependientes de un dominio específico.

Popescu y Etzioni [91] propusieron un sistema de extracción de información de dominio independiente. Identificaron cuatro tareas principales en el análisis de opiniones:

- Identificación de las características del producto;
- Identificación de opiniones con respecto a las características del producto;
- Determinación de la polaridad de la opinión;
- Clasificación de la opinión según su fuerza.

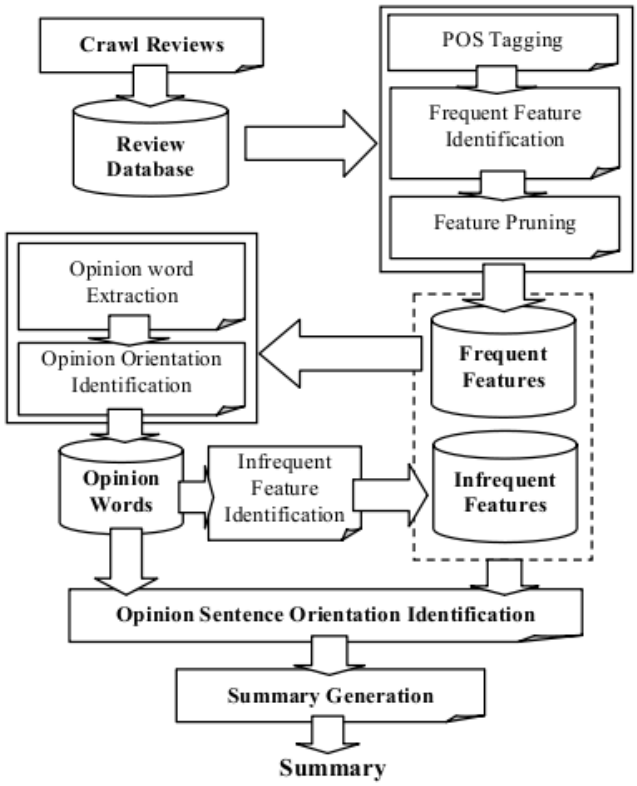

Figura 4.4: Diseño del modelo para minería de opiniones basado en características de Hu y Liu

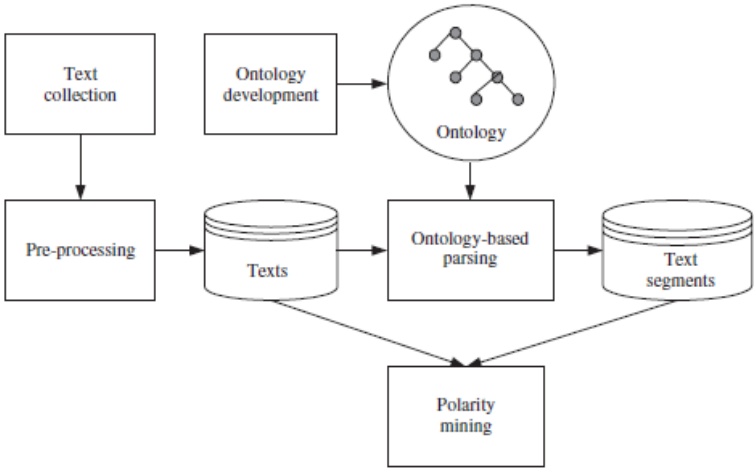

Figura 4.5: Arquitectura para minería de opiniones basado en ontologías de Zhou y Chaovalit

El modelo de Popescu y Etzioni extrae las características de los productos utilizando información mutua (PMI, por sus siglas en inglés). Éste utiliza características explícitas para identificar posibles opiniones basadas en la intuición de que una opinión relacionada con una característica del producto tendrá lugar en sus proximidades del árbol de análisis sintáctico. Después, de la extracción de la opinión, se utiliza la relajación de etiquetado [47] que es una técnica de clasificación no supervisada, para eliminar la ambigüedad de la orientación semántica de las palabras de opinión.

Recientemente se han realizado dos trabajos con el objetivo de calcular la polaridad por medio de las propiedades de un objeto, pero a diferencia del trabajo de Hu y Liu [67] y del trabajo de Popescu y Etzioni [91], en estos trabajos se utilizan ontologías del dominio al que pertenecen los textos del corpus y, a partir de éstas, se extraen los calificativos de los conceptos de los productos que aparecen en los textos.

El primer trabajo de ellos es el estudio de Zhou y Chaovalit en [131] (véase figura 4.5). En dicho estudio el primer paso que se realiza es generar una ontología; la generación de esta ontología se realiza manualmente por analistas a partir del corpus que posteriormente se utilizará para los experimentos de minería de opiniones. Una vez construída la ontología, suponen que dado un concepto c de la ontología, el cual está descrito con n propiedades: $p_1, p_2, ..., p_n$, y que en un texto t se puede expresar opiniones sobre cualquiera de estas propiedades; por tanto, de acuerdo con el número de propiedades de c, t puede dividirse en m segmentos (siendo $m < n$). Es decir que, la predicción de la polaridad de t se define como:

$$polaridad(t) = \begin{cases} positivo, & \text{si } \left(\frac{1}{m}\sum_{i=1}^{n} w_i v_i \geq 0\right) \\ negativo, & \text{en caso contrario} \end{cases} \tag{4.1}$$

siendo $w_i \in [0,1]$ el peso de la propiedad p_i y $v_i \in [-1,1]$ el valor de la polaridad de la propiedad p_i calculadas por estimación de máxima verosimilitud.

El otro estudio que se ha realizado hasta ahora sobre análisis de opiniones basado en ontologías, es el trabajo de Zhao y Li en [130]. En este estudio, se genera una ontología automáticamente a partir del corpus que posteriormente se utilizará para el análisis de opiniones. Una vez generada la ontología, los autores proponen extraer los calificativos de las propiedades de los conceptos por medio de la ontología, para posteriormente identificar la polaridad de dichos calificativos utilizando la herramienta SentiWordNet. Una vez extraídos los calificativos y calculado sus polaridades, obtienen la orientación semántica del texto a partir de la jerarquía de la ontología, para ello calculan la orientación negativa, positiva y neutra según las siguientes ecuaciones:

$$op_{hlc}(neg) = \frac{\sum_{ch\_node_{ws_i \in neg}} score(ch\_node_{ws_i \in neg})}{|ch\_node_{ws_i \in neg}|} \tag{4.2}$$

$$op_{hlc}(pos) = \frac{\sum_{ch\_node_{ws_i \in pos}} score(ch\_node_{ws_i \in pos})}{|ch\_node_{ws_i \in pos}|} \tag{4.3}$$

$$op_{hlc}(neu) = \frac{\sum_{ch\_node_{ws_i \in ne}} score(ch\_node_{ws_i \in ne})}{|ch\_node_{ws_i \in ne}|} \tag{4.4}$$

donde $|ch\_node_{ws_i}|$ representa la cardinalidad de todos los hijos con la misma opinión. Por último, escogen como orientación del texto aquella de las tres que es mayor que el resto.

Sin embargo, en los estudios realizados sobre minería de opiniones, las ontologías no han sido utilizadas únicamente para extraer las características de los productos de los que se opinan. Dey y Haque [15] proponen un modelo que se centra en la extracción de opiniones a partir de documentos de texto. Dey y Haque argumentan que la mayoría de las técnicas existentes de Procesamiento del Lenguaje Natural (NLP, por sus siglas en inglés) asumen que los datos están limpios y correctos. Pero en general las opiniones expresadas en las redes sociales, como comentarios de blogs o las opiniones expresadas por escrito, están llenas de faltas de ortografía y errores gramaticales al estar escritas de manera informal. Su sistema propuesto utiliza una ontología del dominio para extraer las opiniones de los sitios web predefinidos que permite opinar en múltiples niveles de granularidad. Ellos proponen un mecanismo de pre-procesado de texto que explota el conocimiento del dominio para limpiar el texto. Posteriormente, estos textos "limpios" son procesados por las herramientas de PLN para obtener la polaridad de las opiniones. Sin embargo, este proceso es iterativo y difícil de aplicar.

4.3. Análisis de opiniones vía fusión de ontologías

Sin embargo, cuando dos empresas tengan la necesidad de compartir información sobre las opiniones de los productos (o llegado el caso en que dos empresas se fusionen), hay que

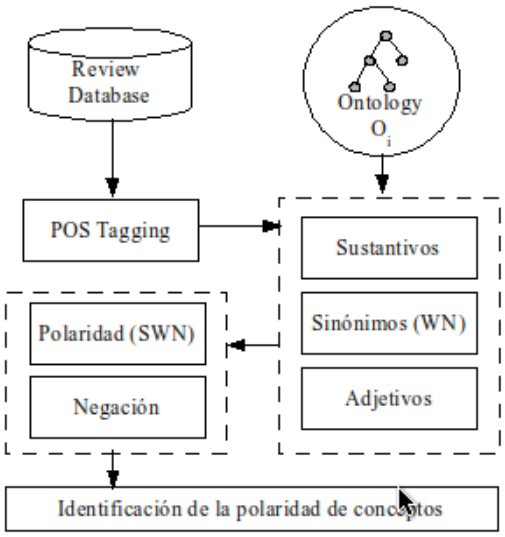

Figura 4.6: Algoritmo para la identificación de la polaridad de conceptos

encontrar una manera para poder analizar las opiniones y posteriormente enviar la información a ambas empresas intentando perder lo menos posible de ésta. En los trabajos realizados hasta la actualidad sobre análisis de opiniones ninguno de ellos ha tenido en cuenta este factor. Para dicho problema nosotros proponemos un algoritmo de análisis de opiniones via fusión de ontologías. Se puede ver un esquema de dicho algoritmo en la figura 4.7. El algoritmo propone que la empresa e_1 obtenga la polaridad de los conceptos de su ontología O_1, del mismo modo la empresa e_2 obtendrá la polaridad de los conceptos de su ontología O_2.

Para la obtención de la polaridad de los conceptos y propiedades de las ontologías se propone los siguientes pasos (se puede ver un esquema en la figura 4.6):

- En el primer paso etiquetamos cada una de las palabras de los textos (Part Of Speech tagger); para este paso hemos utilizado el toolkit GATE[9];
- Posteriormente, buscamos las frases que contienen algún concepto c de la ontología O_i; para ello buscamos en cada texto los nombres (o grupos de nombres) que coinciden con un concepto de la ontología. Para aquellas frases que no contengan ningún concepto de la ontología, utilizamos WordNet para encontrar sinónimos de los nombres que aparecen en la frase, que pueden ser sinónimos a su vez, de algún concepto de la ontología;
- Seguidamente, extraemos de las frases obtenidas en el paso anterior, los adjetivos adyacentes de cada concepto;
- En el siguiente paso obtenemos la polaridad de los adjetivos utilizando SentiWordNet;
- Comprobamos que la frase es afirmativa, en caso contrario, invertimos la polaridad que nos devuelve SentiWordNet;

[9]http://gate.ac.uk/

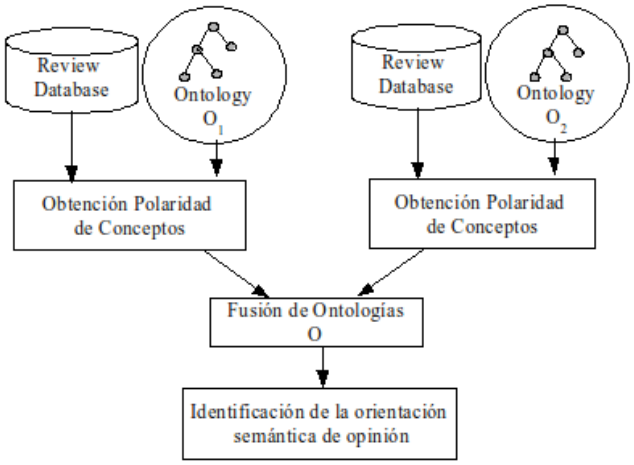

Figura 4.7: Algoritmo para el análisis de opiniones via fusión de ontologías

Posteriormente se realizará una fusión de ontologías mediante una ontología general u ontología superior O (*upper ontology*) y a través de ésta, se realizará un cálculo de la orientación semántica de la opinión t mediante la ecuación:

$$orien\_seman_{neg}(t) = \sum_{c \in O} v_{neg}(c) \tag{4.5}$$

$$orien\_seman_{pos}(t) = \sum_{c \in O} v_{pos}(c) \tag{4.6}$$

donde c son los conceptos que pertenecen a la ontología O, $v_{neg}(c)$ es la polaridad negativa del concepto c y $v_{pos}(c)$ la polaridad positiva. Por tanto, la orientación semántica de la opinión (t) se define como:

$$polaridad(t) = \begin{cases} positiva & \text{si } orien\_seman_{pos}(t) > orien\_seman_{neg}(t) \\ negativa & \text{en caso contrario} \end{cases} \tag{4.7}$$

Hay que destacar, como ya hemos comentado, tanto en el estudio de Zhou y Chaovalit como en el de Zhao y Li, la ontología utilizada se crea a partir del corpus que posteriormente se utilizará en los experimentos para el análisis de opiniones, por otro lado en nuestro trabajo utilizamos ontologías existentes con anterioridad. Nuestra opinión es que esta forma es la más cercana al problema del mundo real, puesto que las empresas tienen una ontología propia.

El primer paso que hemos de hacer para la realización de nuestro algoritmo es encontrar un método de fusión de ontologías que sea eficaz.

4.3.1. Estado del arte

Según Euzenat y Shvaiko [27] un proceso de fusión de ontologías puede ser visto como una función f la cual dadas dos ontologías o y o', un conjunto de parámetros p y un conjunto de oráculos y recursos r, devuelve un alinemainto a entre o y o'.

El problema de la fusión entre ontologías puede ser abordado desde diversos puntos de vista y este hecho se refleja en la variedad de métodos de fusión que se han propuesto en la literatura. Muchos de ellos tienen sus raíces en el problema clásico de la fusión de esquemas en el área de las bases de datos, tales como Artemis [10], COMA [17], Cupid [71], Similarity Flooding [80]...; mientras que otros han sido específicamente diseñados para trabajar con ontologías, como son GLUE [19], QOM [23], OLA [26], S-Match [35], ASCO [61], PROMPT [87], HCONE-merge [58] y SAMBO [59]. Algunos de estos métodos se basan en el razonamiento formal de la estructura de las descripciones de la entidad como S-Match y OLA; otros en cambio, utilizan una combinación de similitud y grafos de razonamiento como en Similarity Flooding; e incluso otros se basan en algoritmos de aprendizaje automático como GLUE.

Los diferentes métodos de fusión de ontologías se pueden dividir en dos grupos [68]:

- **Métodos basados en cadenas**. Estos métodos miden la similitud de ambas entidades sólamente teniendo en cuenta las cadenas de caracteres que forman su nombre. Entre estas medidas se encuentran: SMOA [110] mide la similitud de dos cadenas en función de sus puntos en común en términos de sub-cadenas así como de sus diferencias; distancia de Levensthein [65] mide el número mínimo de inserciones, eliminaciones y sustituciones de caracteres necesarios para transformar una cadena en otra; distancia de n-gramas en la cual dos cadenas de caracteres son más similares cuanto mayor número de n-gramas en común tengan.

- **Métodos basados en lingüística**. Estos métodos explotan las técnicas de PLN para encontrar la similitud entre dos cadenas de caracteres vistos como partes significativas del texto en lugar de secuencias de caracteres. En estos métodos se utilizan también recursos lingüísticos, como tesauros de conocimiento común o de dominio específico, para fusionar palabras a partir de sinónimos, hipónimos, etc.

A continuación describimos algunas de las herramientas que tienen como objetivo la fusión de ontologías:

GLUE [19]: es una versión desarrollada del LSD [18], cuyo objetivo es encontrar semiautomáticamente correspondencias entre esquemas para la integración de datos. GLUE explota técnicas de aprendizaje automático para encontrar correspondencias semánticas entre los conceptos almacenados en diferentes ontologías autónomas [20]. Dadas dos ontologías distintas, el proceso de detección de correspondencia entre sus conceptos se basa en la medida de similitud que se define a través de la distribución de probabilidad conjunta. La medida de similitud entre dos conceptos se calcula como la probabilidad de que una instancia pertenece

a los dos conceptos. La similitud de dos instancias se basa en las reglas aprendidas por el sistema y posteriormente, se utiliza para encontrar la similitud entre el concepto y las relaciones. La correspondencia asignada a una entidad se ve influenciada por las características de sus vecinos del grafo. Las características consideradas se agrupan en particiones con el fin de ser procesadas una sola vez. La probabilidad resultante está compuesta por el análisis de instancias y similitudes de características que son revisadas en múltiple ocasiones. En el conjunto de todas las probabilidades de correspondencia entre los elementos de la ontología, el sistema extrae la correspondencia final mediante la selección de la máxima correspondencia probable.

OLA [26]: OLA (*OWL Lite Alineación*) es una herramienta de alineamiento de ontologías, la cual tiene como características:

- cubre todas las posibles características de ontologías (es decir, terminológicas, estructurales y extensión);
- tiene en cuenta las estructuras de colección (listas, conjuntos) y los tiene en consideración durante la fusión;
- explica las relaciones recursivas y encuentra el mejor alineamiento a través de iteraciones.

OLA tiene como objetivo realizar el proceso automático al nivel máximo posible. Para fusionar las ontologías se basa en una comparación uniforme entre las entidades que componen las ontologías. Las entidades se dividen en categorías (por ejemplo, clases, objetos, propiedades, relaciones) y sólo los elementos de la misma categoría se comparan. Las ontologías son representadas por medio de un grafo, llamado *OL-Graph*, donde los nodos representan las entidades de la ontología, mientras que las aristas representan las relaciones, que son la especialización, la instanciación, la atribución, la restricción y la valoración. La herramienta OLA asigna una función de similitud específica a cada nodo del grafo. La solución del sistema se aproxima por medio de un método iterativo que parte de una similitud local, que se calcula sin tener en cuenta los nodos vecinos de una determinada entidad y, a continuación se integra con las similitudes vecinas. El proceso tiene una cota superior, ya que ninguna de las funciones de similitud puede producir un valor mayor que 1.

S-Match [35]: S-Match es un sistema de fusión de esquemas/ontologías. Se utiliza como datos de entrada dos estructuras en forma de grafo (por ejemplo, los esquemas de bases de datos u ontologías) y devuelve las relaciones semánticas entre los nodos de los grafos, que se corresponden semánticamente entre sí.

S-Match fue diseñado y desarrollado como una plataforma para la fusión semántica, es decir, un sistema altamente modular con un núcleo de computación de relaciones semánticas donde los componentes individuales se pueden conectar, desconectar o personalizar. La arquitectura lógica del sistema se muestra en la figura 4.8.

Los esquemas de entrada (*input schemas*) están codificados en formato XML interno. El módulo encargado de introducir los esquemas/ontologías de entrada se encarga también del

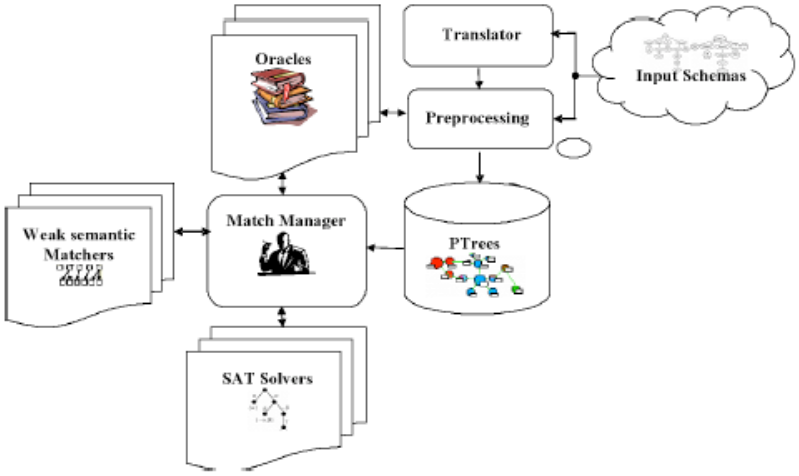

Figura 4.8: Arquitectura de la plataforma S-Match

preprocesamiento. En particular, se calcula descendentemente para cada etiqueta de un árbol, el sentido capturado por las etiquetas dadas en un esquema u ontología utilizando las técnicas descritas en [72]. El módulo de preprocesamiento (*Preprocessing*) tiene acceso al conjunto de oráculos (*Oracles*) que proporcionan el conocimiento léxico y del dominio necesario a priori. La salida del módulo es un árbol enriquecido. Estos árboles enriquecidos se almacenan en una base de datos interna (*PTree*) donde se puede navegar, editar y manipular.

El administrador de fusión (*match manager*) coordina el proceso de fusión con tres librerías extensibles. La primera librería está compuesta de lo que se denomina en [35], fusionadores de semántica débil a nivel de elemento (*weak semantics element level matchers*). Estos realizan manipulaciones de cadenas (por ejemplo, prefijo, análisis de n-gramas, distancia de edición, tipos de datos, etc) y tratan de adivinar la relación semántica implícita codificada en palabras similares. La segunda librería está formada por fusionadores de semántica fuerte a nivel de elemento, llamadas oráculos (*Oracle*). La tercera librería está formada por fusionadores de semántica fuerte a nivel de estructura, llamados *SAT solvers*.

ASCO [61]: ASCO adopta un algoritmo que identifica los pares de elementos correspondientes de dos ontologías diferentes. Estos pares pueden ser pares de conceptos (clases) en las dos ontologías o pares de relaciones, o incluso parejas de un concepto en una ontología y una relación en la otra ontología. El proceso de comparación de ASCO se compone de varias fases. La fase lingüística aplica técnicas de procesamiento lingüístico, y utiliza métricas de comparación de cadenas, y bases de datos léxicas para calcular la similitud entre dos conceptos o entre dos relaciones. Durante la fase de procesamiento lingüístico, ASCO en primer lugar normaliza los términos, a partir de signos de puntuacion, mayúsculas, símbolos especiales, dígitos para tener un conjunto de *tokens*. Estos *tokens* son comparados por medidas de comparación de cadenas como Jaro-Winkler [124], Levenstein [65] o Monger-Elkan [82]. También se integra WordNet para aumentar la precisión y para evitar los problemas de los conflictos de términos. Las similitudes lingüísticas calculadas son los datos de entrada para

la fase estructural. En esta fase, ASCO trata de explotar la estructura de la ontología para modificar o consolidar la similitud de los dos conceptos o las dos relaciones. Las similitudes de clases o de relaciones son iterativamente transmitidos a sus vecinos en el árbol de la ontología que se construye a partir de la jerarquía de clases y la jerarquía de las relaciones. Cuando termina la propagación, si las similitudes entre las clases o las relaciones superan un umbral, se consideran como similares.

HCONE-merge [58]: Su contribución a la fusión de ontologías es una forma de alineación de conceptos especificado en una ontología para los sentidos de palabras utilizando Análisis Semántico Latente (LSA, por sus siglas en inglés). Los humanos están involucrados en el proceso de fusión en dos etapas: en la captación de la semántica de los términos previstos por medio de definiciones informales, con el apoyo de Indexación Semántica Latente (LSI, por sus siglas en inglés), y en el esclarecimiento de las relaciones entre los conceptos en caso de que tales relaciones no se declaren formalmente. LSI es una técnica de espacio vectorial para la recuperación de la información y la indexación. Básicamente, para un concepto dado C, encuentra los sentidos de la palabra lexicalizada por C o sus variaciones buscando en WordNet, y amplía los sentidos de la palabra por hiponimia. El espacio semántico para C está formado por una matriz de $n * m$ que comprende los n términos vecinos de mayor frecuencia de los m sentidos de la palabra. LSA se utiliza para encontrar el mejor sentido de la palabra asociada con una cadena de consulta usando los términos vecinos de C. Sin embargo, sólo las subclases y superclases están incluidas en la vecindad del concepto, y las descripciones del concepto no se consideran en la cadena de consulta.

4.3.2. Ontologías superiores (*upper ontologies*)

Una *upper ontology*, como se define en [90], es una ontología independiente del dominio, que proporciona un marco por el cual distintos sistemas pueden utilizar una base común de conocimiento y desde el cual se pueden derivar ontologías de dominio específico. Los conceptos expresados son destinados a ser fundamentales y universales para garantizar la generalidad y la expresividad de una amplia gama de dominios [100]. Una *upper ontology* se caracteriza a menudo como la representación de conceptos que son básicos para la comprensión humana del mundo [56]. Por lo tanto, una *upper ontology* se limita a los conceptos que son genéricos, abstractos y filosóficos.

Existen varias *upper ontologies* implementadas, como BFO [36], Cyc [64], DOLCE [32], GFO [41], PROTON [9], Sowa's ontology [104] y SUMO [86]. Se puede encontrar una comparación de las distintas *upper ontologies* mencionadas anteriormente en [76].

En nuestros experimentos hemos utilizado SUMO y OpenCyc. En los siguientes subapartados describimos cada una de estas ontologías generales.

SUMO

SUMO[10] (Suggested Upper Merged Ontology) es una ontología creada por Teknowledge Corporation[11] con una amplia contribución de la lista de correo SUO[12] (Standard Upper Ontology), y fue propuesta como documento de iniciación para el Grupo de Trabajo SUO [86], un grupo de trabajo con colaboradores de los campos de la ingeniería, la filosofía y ciencias de la información. SUMO es una de las más grandes ontologías formales públicas existentes hoy en día. Cuenta con 20.000 términos y 70.000 axiomas cuando todos los dominios son combinados. SUMO está compuesto por una ontología de nivel medio (MId-Level Ontology (MILO)), ontologías de comunicaciones, países y regiones, computación distribuida, economía, finanzas, componentes de ingeniería, geografía, gobierno, militar, sistema de clasificación industrial de Norte América, gente, los elementos físicos, cuestiones internacionales, transporte, aeropuertos mundiales y armas de destrucción masiva.

OpenCyc

El Cyc Knowledge Base[13] (KB) es una base de conocimiento multicontextual desarrollada por Cycorp. Cyc es una representación formalizada de una cantidad enorme de conocimiento fundamental humano: hechos, reglas básicas, y heurísticas para razonar sobre los objetos y los acontecimientos de la vida cotidiana. Cyc KB consiste en términos y las aserciones que relacionan los términos.

KB Cyc se divide en varias *microteorías*, cada una es esencialmente un conjunto de las aserciones que comparten un juego común de suposiciones; algunas *microteorías* son enfocadas en un dominio particular de conocimiento, en un nivel particular de detalle, etc.

Actualmente, el KB Cyc contiene casi trescientos mil términos y cerca de tres millones de aserciones (hechos y reglas) utilizando más de quince mil relaciones. Cyc es un producto comercial, sin embargo Cycorp dispone de una versión de código abierto OpenCyc[14], y una versión para uso de investigación ResearchCyc[15].

4.3.3. Fusión de ontologías vía *upper ontology*

Existen estudios enfocados a la fusión de ontologías mediante la utilización de ontologías generales como fondo. El primer trabajo que conocemos es el trabajo de Li [66], en el que desarrolla la herramienta LOM (*Lexicon-based Ontology Mapping Tool*). LOM es un prototipo de herramienta de correspondencias de ontologías basada en el léxico desarrollado por

[10] http://dream.inf.ed.ac.uk/projects/dor/sumo/
[11] http://www.teknowledge.com/
[12] http://suo.ieee.org/
[13] http://www.cyc.com
[14] http://opencyc.org/
[15] http://researchcyc.cyc.com/

Teknowledge en 2004. Utiliza cuatro métodos para fusionar los vocabularios de cualquiera de las dos ontologías: fusionar todo los términos, fusionar palabras constituyentes, fusionar *synset*, y fusionar tipos.

El método para fusionar tipos explota las correspondencias entre los synsets de WordNet y las ontologías SUMO y MILO. Al utilizar las correspondencias de SUMO-WordNet, LOM ayuda a resolver un problema evidente en la fusión de términos, proporcionando una gran cantidad de sinónimos. LOM etiqueta cualquier sinónimo de un concepto c en la ontología o con el concepto de SUMO, y hace lo mismo con los conceptos de o'. Luego, compara estas etiquetas.

En [4], Aleksovski y sus colaboradores discuten los experimentos realizados para fusionar porciones de CRISP[16] y MeSH[17] utilizando como ontología de fondo FMA[18]. En estos experimentos, Aleksovski y sus colaboradores demostraron que se obtenían mejores resultados en la fusión utilizando FMA como ontología de fondo, que realizando la fusión directamente entre las dos ontologías.

Sin embargo, en ninguno de estos trabajo se explota el uso de las ontologías superiores (*upper ontologies*). En [78] se describe un trabajo preliminar sobre la fusión de ontologías vía ontologías superiores. En [77] Locoro y sus colaboradores describen un algoritmo para la fusión de ontologías vía ontologías superiores. Este algoritmo consiste en fusionar en paralelo la ontología o con la ontología superior u devolviendo una alineación a, y al mismo tiempo fusionar la ontología o' con la ontología u devolviendo la alineación a'. Posteriormente, realiza una composición entre las alineaciones devueltas a y a' obteniendo la alineación final a''. Locoro *et al.* [68] implementaron dos diferentes métodos para la composición entre alineaciones:

- **El método no estructurado**: En la etapa de composición entre alineaciones las entradas son: a que corresponde con el alineamiento entre la ontología o y la ontología superior u, y a' que corresponde con el alineamiento entre la ontología o' y u. Si el concepto $c \in o$ y $c' \in o'$ corresponden al mismo concepto $c_u \in u$ entonces c y c' están relacionados, con una medida de confianza de $conf_1 * conf_2$ siendo $conf_1$ la medida de confianza del concepto c con el concepto c_u en el alineamiento a y $conf_2$ la medida de confianza del concepto c' con el concepto c_u en el alineamiento a'. En la figura 4.9 puede verse un esquema del método no estructurado.
- **El método estructurado**: En la etapa de composición entre alineaciones las entradas son: a que corresponde con el alineamiento entre la ontología o y la ontología superior u; a' que corresponde con el alineamiento entre la ontología o' y u; y un factor de decadencia df. Por tanto el concepto $c \in o$ y $c' \in o'$ están relacionados si:
  - los conceptos c y c' corresponden al mismo concepto $c_u \in u$, por tanto están relacionados con una medida de confianza de $conf_1 * conf_2$.

[16]The Computer Retrieval of Information on Scientific Projects: http://crisp.cit.nih.gov/
[17]The Medical Subject Headings: http://www.nlm.nih.gov/mesh/
[18]The Foundation Model of Anatomy ontology: http://sig.biostr.washington.edu/projects/fm/

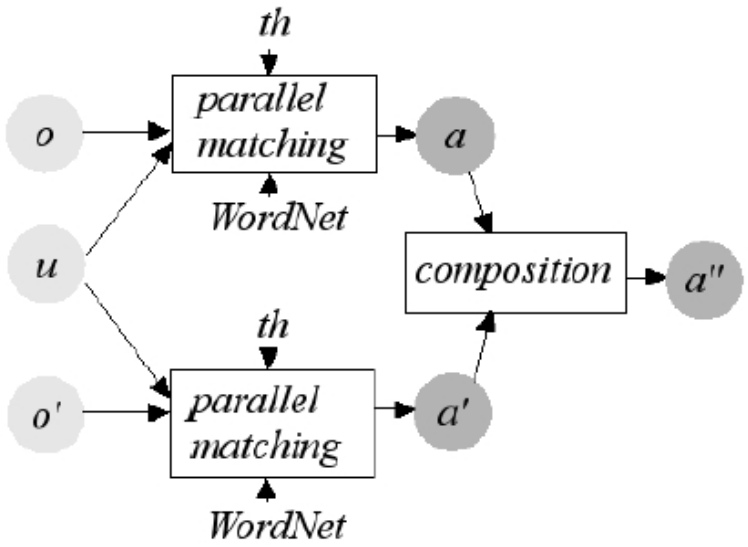

Figura 4.9: Esquema para la fusión de ontologías vía ontologías superiores con el método no estructurado

- el concepto c corresponde al concepto c_u, c' corresponde con el concepto c'_u, y c'_u es un super-concepto de c_u en u (o viceversa), por tanto están relacionados con una medida de confianza de $conf_1 * conf_2 * df$.
- el concepto c corresponde al concepto c_u, c' corresponde con el concepto c'_u, y c'_u tiene algún super-concepto en común con c_u, por tanto están relacionados con una medida de confianza de $conf_1 * conf_2 * df^2$.

En la figura 4.10 puede verse un esquema del método estructurado.

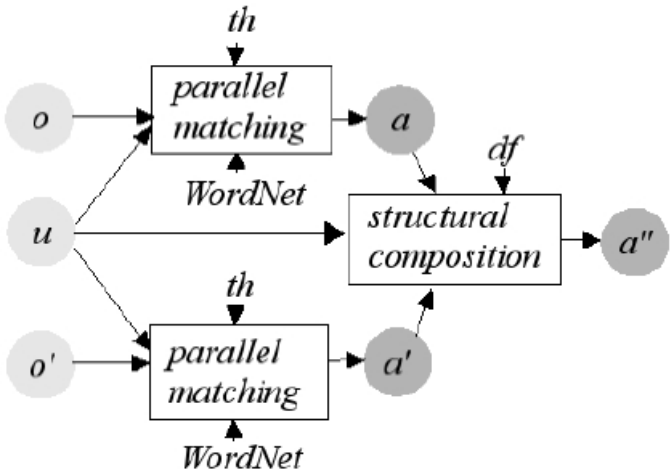

Figura 4.10: Esquema para la fusión de ontologías vía ontologías superiores con el método estructurado

Capítulo 5

Evaluación

La intención de este capítulo es la de analizar los experimentos que hemos realizado para validar el algoritmo propuesto de análisis de opiniones vía fusión de ontologías en el dominio del turismo. Los experimentos se han divido en dos fases diferentes: en la primera fase hemos realizado un estudio para encontrar el método más idóneo para la etapa de fusión de ontologías; y en la segunda fase, hemos realizado los experimentos dedicados a la validación del algoritmo propuesto de análisis de opiniones vía fusión de ontologías en el dominio del turismo.

Este capítulo está organizado de la siguiente manera: en la sección 5.1 comentamos los experimentos que hemos realizado en la primera fase y en la sección 5.2 analizamos las pruebas realizadas en la segunda fase.

5.1. Fusión de ontologías de turismo vía *upper ontologies*

Como ya hemos comentado, el algoritmo que hemos propuesto para el análisis de opiniones vía fusión de ontologías contiene una fase de fusión de ontologías, por tanto, el primer paso era decidir qué técnica de fusión de ontologías es más idónea para nuestro dominio. Con este propósito hemos realizado diferentes experimentos en los que hemos analizado técnicas directas con diferentes medidas de similitud, y técnicas vía *upper ontology*.

5.1.1. Corpus utilizado

Para la realización de los diferentes experimentos hemos decidido utilizar el turismo como dominio de las diferentes ontologías. Hemos utilizado Swoogle[1] para la búsqueda de ontologías

[1] http://swoogle.umbc.edu/

del dominio turismo, en el que aparecían un total de 96 ontologías. Sin embargo, de los 96 ontologías no todas cumplían las expectativas, ya que existían ontologías repetidas, otras estaban en diferente idioma al inglés, o pertenecían a un dominio diferente como puede ser la economía. En la tabla 5.1 aparece detallado este aspecto.

Resultado de la búsqueda	96
Ontologías no encontradas	29
No del dominio turismo	25
Repetidas	21
Otro Idioma	16
Ontologías reales	5

Tabla 5.1: Detalle de ontologías encontradas

En definitiva, como se observa en la tabla 5.1, se contaba con un total de cinco ontologías diferentes. En la tabla 5.2 mostramos las ontologías que hemos utilizado junto con la cantidad de conceptos diferentes que tiene cada una de ellas. Es cierto que en estas ontologías el número de conceptos es muy bajo; no obstante, son las únicas ontologías reales del dominio del turismo que están disponibles en la Web, por lo que hemos decidido utilizarlas para los experimentos.

Ontologías	**Conceptos**
http://www.info.uqam.ca/Members/valtchev_p/mbox/ETP-tourism.owl	194
http://qallme.itc.it/ontology/qallme-tourism.owl	125
http://homepages.cwi.nl/∼troncy/DOE/eon2003/Tourism-ProtegeExportOWL.owl	86
http://fivo.cyf-kr.edu.pl/ontologies/test/VOTours/TravelOntology.owl	35
http://e-tourism.deri.at/ont/e-tourism.owl	20

Tabla 5.2: Ontologías utilizadas en los experimentos

Para las pruebas de los métodos de fusión de ontologías vía *upper ontology* hemos utilizado como ontologías generales, SUMO y OpenCyc[2].

5.1.2. Medidas de evaluación

Para poder medir la eficacia de las diferentes técnicas de fusión de ontologías, necesitábamos encontrar el alineamiento óptimo entre los conceptos de cada ontología. Este alineamiento nos servía como base para medir el alineamiento que nos devolverá cada experimento. Este alineamiento base no podía ser creado automáticamente (por razones obvias, ya que nuestro objetivo es encontrar un método automático capaz de devolver el alineamiento más próximo a éste), así que creamos manualmente un alineamiento entre todos los posibles pares de ontologías.

[2]En el capítulo 4 sección 4.3 describimos ambas ontologías generales

Para comprobar la eficacia de los métodos de fusión de ontologías hemos utilizado las medidas de precisión, recall y F-measure. Dado R el alineamiento base, entonces $|R|$ corresponde al número de alineaciones entre conceptos en el alineamiento R; dado S el alineamiento devuelto por la fusión de ontologías, entonces $|S|$ corresponde al número de alineaciones entre conceptos en el alineamiento S; dado S_R el conjunto de alineaciones de S que coincide en R, entonces $|S_R|$ corresponde al número de alineaciones entre conceptos en el alineamiento S_R. Las fórmulas para calcular las medidas de precisión y recall son:

$$precision = \frac{|S_R|}{|S|} \tag{5.1}$$

$$recall = \frac{|S_R|}{|R|} \tag{5.2}$$

5.1.3. Discusión de los resultados

En primer lugar hemos realizado los experimentos sobre técnicas de fusión de ontologías directas. Para estas pruebas hemos utilizado la API de Euzenat[3]. La API de Euzenat o Alignment API [25] desarrollada por Jérôme Euzenat es una API para la ejecución de alineamientos entre ontologías. El Alignment API esta implementada en Java y permite la ejecución de diferentes medidas de distancia entre textos. Para los test hemos utilizado las medidas *equal* que compara si el texto es exacto, *SMOA* (String Metric for Ontology Alignment) [110] y por último *Levenshtein*[4] [65].

Una vez terminadas todas las pruebas con las técnicas directas, hemos realizado las pruebas con las técnicas vía *upper ontology*. Para ello hemos utilizado la API desarrollada por Angela Locoro et al [77]. El API está desarrollada en Java y permite la ejecución de métodos estructurados y no estructurados[5]. En la ejecución primero hemos utilizado la API de Euzenat entre la *upper ontology* y cada una de las ontologías a alinear. Al utilizar la API de Euzenat nos permitía ejecutar cada una de las funciones de distancia anteriormente citadas, de este modo podíamos hacer una mejor comparación entre técnicas directas y vía *upper ontology*. Una vez realizados los alineamientos de las ontologías seleccionadas con las *upper ontologies* utilizamos el API de Locoro para realizar los alineamientos utilizando el método estructurado y no estructurado. Este proceso lo hemos realizado para cada una de las dos *upper ontologies*, SUMO y OpenCyc.

Una vez terminadas las pruebas hemos colocado los resultados en tablas para poder comparar los resultados. Los campos que aparece en la tabla corresponden a: función de distancia (Distancia), *upper ontology* utilizada (Upper Onto), método estructurado o no estructurado (Método), número de alineamientos encontrados (Encontrados), número de alineamientos

[3]http://alignapi.gforge.inria.fr/

[4]En el capítulo 4, sección 4.3.1 explicamos en qué consiste el cálculo de estas medidas

[5]En el capítulo 4, sección 4.3.3 explicamos en qué consiste los métodos estructurados y no estructurados

correctos (Corrrec), y las tres medidas: precisión (Prec), recall (Rec) y F-measure (F-Meas). Por último, hemos señalado en la tabla los mejores resultados para cada test en las diferentes medidas.

En el Apéndice B se muestran los resultados obtenidos en cada test de los experimentos. Lo primero que llama la atención de los resultados, son aquellos tests donde como medida de precisión se obtiene un 1 (veáse tablas A.1, A.4, A.5, A.6, A.8 y A.9). Sin embargo, observando el número de conceptos encontrados podemos ver que son muy pequeños entre 1 y 5 conceptos. Además, estos casos se producen la mayoría de veces utilizando las técnicas via *upper ontology* con SUMO y como función de medida de distancia "equal"; como la función "equal" enlaza conceptos con exactamente el mismo texto, esto nos da a entender que la *upper ontology* SUMO tiene pocos términos pertenecientes al turismo.

Siguiendo con el análisis, si observamos los datos podemos ver los siguientes puntos:

- Cuando aplicamos la técnica de *upper ontology* con el método no estructurado obtenemos una pérdida media de precisión frente a los métodos directos, de alrededor del 1 % para SUMO y de un 2.5 % para OpenCyc. En recall se produce una pérdida media de 19.5 % para SUMO y de 24.5 % para OpenCyc. Y finalmente, en F-measure se produce una pérdida media de 26.5 % para SUMO y de 19.9 % para OpenCyc.

- Cuando aplicamos la técnica de *upper ontology* con el método estructurado obtenemos una pérdida media de precisión frente a los métodos directos, de alrededor del 27 % para SUMO y de un 45 % para OpenCyc. En recall se produce una pérdida media de 19.5 % para SUMO y de 35.5 % para OpenCyc. Y finalmente, en F-measure se produce una pérdida media de 26.5 % para SUMO y de 42.3 % para OpenCyc.

Estos datos nos indican que se obtienen mejores resultados con las técnicas directas que realizando la fusión de ontologías vía *upper ontology*. Sin embargo analizando los diferentes tests podemos sacar otras conclusiones.

En los tests donde las dos ontologías son muy diferentes aún siendo del mismo dominio, como es el caso del test 5 (tabla A.5) o el test 9 (tabla A.9), en el que sólo pueden enlazarse 4 términos de cada ontología, se observa como en estos casos se obtienen mejores resultados con los métodos directos, ya que aunque vía *upper ontology* enlace el mismo número de términos correctos, alinea más posibles enlaces de los que realmente son. Esto ocurre porque como las *upper ontology* son ontologías mucho mayores que las ontologías utilizadas para los experimentos, existe una gran probabilidad de que encuentre términos más semejantes.

Por otro lado, en aquellos tests donde el número de conceptos es un poco más alto como los tests 6, 7 y 8 (tablas A.6, A.7 y A.8), se nota una mejoría en la medida de precisión en los métodos vía *upper ontology*, aunque se continúa obteniendo mejores resultados con los métodos directos. La causa de que continúe encontrando más alineamientos que en los métodos directos, continúa siendo la misma, ya que aunque son un poco más grandes la ontologías no son lo suficientemente como para compararse a las *upper ontologies* utilizadas.

En cambio, en recall se aprecia una notable mejoría, incluso llegando a ser mejor los métodos vía *upper ontology*.

Sin embargo, en las pruebas donde aumenta el número de términos de cada ontología como en el test 3 (tabla A.3), es notable una mejora en los métodos utilizando *upper ontology* respecto a los métodos directos en cuanto a las medida de precisión y F-measure, y una mínima pérdida en recall. Esta mejoría se nota sobre todo al utilizar OpenCyc como *upper ontology*. Esto nos puede dar un indicio de que cuanto mayor son las ontologías es preferible utilizar las técnicas vía *upper ontology*.

Siguiendo con el anásis, si comparamos las dos *upper ontologies* que hemos utilizado, SUMO y OpenCyc, se observa que con OpenCyc hay una leve mejora de un 0.82 % y un 0,67 % en precisión y F-measure respectivamente. En cambio, en la medida de recall se observa un 10,69 % de mejoría utilizando OpenCyc. Por tanto, es lógico pensar que es preferible utilizar en el dominio del turismo OpenCyc como *upper ontology*.

Por último, comparando los métodos de vía *upper ontology* estructurado y no estructurado, se observa que con los métodos no estructurados se obtienen una mejora en la precisión de un 27.1 % en el caso de SUMO y de un 47.44 % en OpenCyc. En la medida de recall se obtienen exactamente los mismo resultados en los dos métodos salvo en el test 8 (tabla A.8), donde con OpenCyc se obtiene una leve mejoría utilizando los métodos no estructurados. Por último, en F-measure se obtiene una mejora con los métodos no estructurados de un 16.30 % utilizando SUMO y de un 32.5 % en OpenCyc.

Hay que destacar también, que en el caso de los métodos estructurados y no estructurados el tamaño de las ontologías no importa, es decir, siempre se obtienen mejores resultados utilizando los métodos no estructurados.

Estos datos nos podrían indicar que se obtienen mejores resultados con las técnicas directas que realizando la fusión de ontologías vía *upper ontology*. Sin embargo, en los diferentes experimentos se ha comprobado que conforme aumenta el número de términos de cada ontología, es notable una mejora en los métodos utilizando *upper ontology* respecto a los métodos directos en cuanto a las medida de precisión y F-measure, y una mínima pérdida en recall. Esta mejoría se nota sobre todo al utilizar OpenCyc como *upper ontology*. Este hecho se puede comprobar en la tabla A.3. Esto nos puede dar un indicio de que cuanto mayor son las ontologías es preferible utilizar las técnicas vía *upper ontology*.

En definitiva, analizando los resultados obtenidos, hemos decidido que la técnica para la fase de fusión de ontologías sea la fusión de ontologías vía *upper ontology* con el método no estructurado y utilizando OpenCyc como ontología general.

5.2. Análisis de opiniones vía fusión de ontologías

Una vez decidido la técnica para la fusión de ontologías, hemos realizado los experimentos para el análisis de opiniones vía fusión de ontologías. En los experimentos hemos intentado

simular como actuarían dos empresas dedicadas al dominio del turismo para obtener información sobre algunos de los productos ofertados.

Con este propósito, primero hemos realizado un proceso de análisis de opiniones sobre el corpus con dos diferentes ontologías del dominio del turismo. Posteriormente, suponemos que las empresas desean intercambiar información, compartirla, o en el caso extremo en que dos empresas se fusionen, por tanto realizarán un proceso de fusión de ontologías.

5.2.1. Corpus utilizado

Para la realización de los experimentos hemos creado un corpus formado por 3.000 textos cortos de opiniones, de los cuales 1.500 eran opiniones positivas y 1.500 negativas[6]. Puesto que nuestros experimentos se centran en el análisis de opiniones de productos o servicios, realizadas en redes sociales tales como blogs, hemos utilizado la página web de TripAdvisor[7] para extraer opiniones relacionadas sobre conceptos del dominio del turismo como hoteles, restaurantes y ciudades. Estos textos son de tamaño reducido, siendo el tamaño máximo de 7 KB.

Una ventaja que nos aportaba TripAdvisor, es que en este blog los usuarios no sólo escriben sus opiniones sino que además puntúan el producto entre "Excelente", "Muy bueno", "Regular", "Malo" y "Pobre". A partir de esta puntuación hemos etiquetado los textos, es decir, como textos positivos hemos utilizado las opiniones calificadas como "Excelente" y "Muy Bueno", y para los textos negativos, aquellas con calificaciones "Malo" y "Horrible". De esta forma no teníamos que analizar las opiniones manualmente para clasificar el valor de la orientación semántica de cada una de ellas.

Por último, señalar que para la realización de las pruebas hemos utilizado como ontologías "ETP-Tourism" y "qallme-tourism", ya que son las ontologías con mayor número de conceptos de todas las disponibles.

5.2.2. Medidas de evaluación

El análisis de opiniones consiste en una tarea de clasificación de la orientación semántica de la opinión en dos clases: positiva o negativa. Por tanto, para medir la eficacia de nuestro algoritmo hemos tenido en cuenta el porcentaje de aciertos en la clasificación de las diferentes clases.

[6]En http://users.dsic.upv.es/grupos/nle/?file=kop4.php se puede obtener el corpus
[7]http://www.tripadvisor.com

Ontología	Num.	Adj.	Adj. + Vb	Adj. + Adv.	Adj. + Vb. + Adv.
ETP Tourism	1.500	72,41 %	72,16 %	69,47 %	68,93 %
qallme-tourism	1.500	70,92 %	71,2 %	68,21 %	67,93 %
Ontology matching	3.000	71,13 %	71,56 %	68,88 %	68,63 %

Tabla 5.3: Resultados obtenidos dividiendo el corpus

5.2.3. Discusión de los resultados

En los experimentos hemos realizado distintas pruebas utilizando, en cada uno de éstas, diferentes palabras para el cálculo de la polaridad para comprobar si además de los adjetivos (ya que son las palabras utilizadas para calificar los conceptos, además de ser los adjetivos las palabras más utilizadas en la mayoría de trabajos realizados hasta la fecha), podía mejorarse los resultados añadiendo palabras de otra categoría, como por ejemplo, verbos. Las posibilidades en las que hemos experimentado son: solo adjetivos (Adj.), adjetivos y verbos (Adj. + Vb.), adjetivos y adverbios (Adj. + Adv.) y adjetivos, verbos y adverbios (Adj. + Vb. + Adv.).

Para poder medir mejor la eficacia del algoritmo propuesto, hemos realizado dos diferentes experimentos: en el primer experimento hemos separado el corpus para cada una de las dos empresas, con la intención de simular que ocurriría si dos empresas analizan diferentes textos antes de compartir la información sobre el análisis de opiniones; y en el segundo, hemos utilizado el corpus completo para las dos ontologías, simulando que dos empresas analizan anteriormente los mismos textos.

En la tabla 5.3 se muestran los resultados obtenidos en los experimentos dividiendo el corpus en dos. Un dato destacable es que tras realizar el proceso de fusión de ontologías se obtienen siempre resultados muy cercanos a los resultados obtenidos con los obtenidos por separado en cada ontología, es más, aunque los resultados son un poco inferiores comparándolo con los resultados obtenidos con la ontología *ETP-Tourism*, son un poco superiores que con la ontología *qallme-tourism*.

Los resultados del segundo experimento en el que utilizamos el corpus completo en las pruebas con cada ontología, se muestra en la tabla 5.4. Como se observa en la tabla, el resultado tras realizar el proceso de fusión de ontologías es muy similar al obtenido en el experimento anterior. Estos resultados nos dan a entender que al realizar el proceso de fusión de ontologías no se pierden datos referentes al proceso de análisis de opiniones realizado con antelación.

Por otra parte, como se comprueba, tanto en la tabla 5.3 como en la tabla 5.4, incluyendo los verbos a los adjetivos (Adj. + Vb.) para calcular la polaridad se obtiene una leve mejora con respecto a calcular la polaridad únicamente a través de los adjetivos (Adj.). También se observa que con esta opción (Adj. + Vb.), los resultados individuales en cada una de las ontologías, son inferiores que utilizando la opción de únicamente adjetivos (Adj.). Sin

Ontología	Num.	Adj.	Adj. + Vb	Adj. + Adv.	Adj. + Vb. + Adv.
ETP Tourism	3.000	72,2 %	71,56 %	69,56 %	69,06 %
qallme-tourism	3.000	71,2 %	71,03 %	68,13 %	68,33 %
Ontology matching	3.000	71,33 %	71,53 %	68,93 %	68,86 %

Tabla 5.4: Resultados obtenidos con el corpus completo

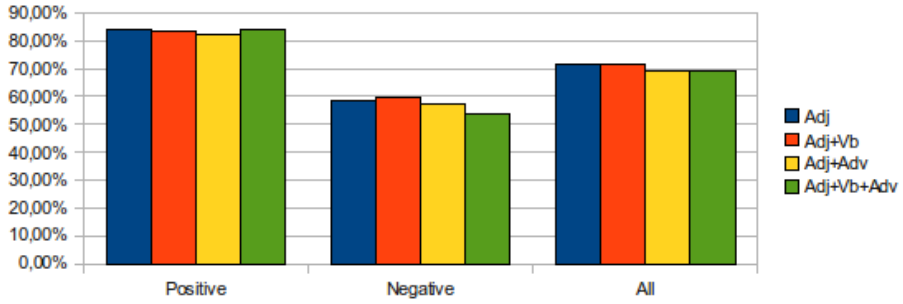

Figura 5.1: Resultados de minería de polaridad

embargo, tras la fusión de ontologías se obtiene un resultado superior con la opción de adjetivos y verbos.

En la figura 5.1 se muestra una comparación entre los resultados obtenidos con cada una de las opciones que hemos experimentado. También hemos incluido una representación de los resultados separando las opiniones positivas de las opiniones negativas. Como se observa, utilizando los adjetivos más verbos para calcular la polaridad se observa que hay una leve mejoría de 0.2 puntos. Pero también hay que destacar que con esta misma opción la diferencia entre el porcetaje de aciertos con las opiniones positivas y el porcentaje de aciertos con las opiniones negativas, es menor que en el resto de opciones.

Una interesante observación que se desprende de los resultados de los experimentos y que se ve reflejado en la figura 5.1 es la diferencia de porcentajes de aciertos que existe cuando examinamos únicamente las opiniones positivas frente a cuando examinamos las opiniones negativas. Analizando los textos que componen nuestro corpus hemos observado que las personas cuando estamos en desacuerdo con algún producto o servicio tenemos cierta tendencia a utilizar la ironía y el sarcasmo [117, 95]. Ésto provoca que al extraer los adjetivos para obtener la orientación semántica, éstos tendrán la polaridad cambiada, llevando a clasificar los textos con la polaridad incorrecta. Pero este hecho, aunque es mucho más frecuente en opiniones negativas, también se utiliza pero con menor medida en opiniones positivas.

Capítulo 6

Conclusiones

Las nuevas tecnologías han evolucionado de una forma increíble en los últimos años. Hace poco veíamos en las películas, agentes secretos con teléfonos móviles y nos reíamos de la imaginación de los guionistas de Hollywood. Sin embargo, como dice el dicho "la realidad supera la ficción"; y es que actualmente los móviles han avanzado tanto que ya no sólo sirven para hablar con los conocidos, sino que se pueden comunicar a través de las redes sociales de la Web 2.0. Es más la tecnología ha avanzado tanto que hoy en día es muy común encontrar a un particular en un bar con un *laptop* diseñando algo en Photoshop o Autocad, o editando algún comentario de alguna red social como Facebook o MySpace.

Nos encontramos en la era de la Web 2.0, la era en la que las empresas deben aprovechar al máximo las oportunidades que ofrece Internet. Según un estudio de la Fundación de la Innovación Bankinter, la Web 2.0 es una web participativa, inteligente y eficaz, por ello las empresas deberían ofrecer un flujo de conocimiento ilimitado que ahorrará tiempo al usuario, consiguiendo nuevas oportunidades de negocio [85].

Este flujo de información es donde se encuentran tanto las mayores ventajas como desventajas que aporta la Web 2.0. Una de las principales preocupaciones de las empresas es la seguridad de la propiedad intelectual, así como la seguridad de los datos confidenciales. No obstante, analizar el flujo de información de los usuarios, los cuales hacen pública su inconformidad o lealtad a determinados productos o servicios siendo compartida dicha información con el resto de usuarios, permite a las empresas responder públicamente aportando información a sus clientes y modificando sus productos o servicios según la tendencia del mercado. En este trabajo hemos analizado estos puntos y hemos llegado a las siguientes conclusiones.

6.1. Cómo protegerse de las desventajas de la Web 2.0

Con la llegada de la Web 2.0 se ha producido un aumento en el número de plagios entre empresas. Esto tiene dos principales explicaciones: la facilidad de acceso a la información que proporciona la Web 2.0 y la cantidad de información, la cual proporciona cierta seguridad al

que comete plagio por la dificultad de ser descubierto.

Sin embargo, plagiar no solamente es copiar un producto o servicio, sino que conlleva el robo del mérito, del prestigio y de cierta parte de la identidad de la empresa. Una empresa debe proteger su material intelectual, pues su mayor éxito en el mercado es su propia identidad, aquellos productos o servicios que la diferencian del resto de empresas y por lo que es conocida.

En este trabajo hemos tratado de ponernos en la piel de una empresa y en su necesidad de detectar los casos de plagio de sus campañas de marketing, sus ideas, etc publicadas en la Web. La idea era investigar la eficacia de las herramientas disponibles, como WCopyFind, que una empresa tiene a su alcance para poder detectar casos de plagio. Los pobres resultados que obtuvimos con la herramienta WCopyFind en la competición PAN de detección de plagio, así como los que se obtieveron con el sistema Ferret, nos han demuestrado la necesidad de desarrollar métodos *ad hoc* de detección automática de plagio para las empresas.

En este trabajo hemos comprobado que a diferencia de la mayoría de las áreas del Lenguaje Natural, en la detección automática de plagio, la precisión es menor que el recall. Esto se debe a que es muy probable encontrar fragmentos similares entre los dos documentos, aunque estos no sean fragmentos plagiados. Para un trabajo futuro, sería interesante la búsqueda de un enfoque automático para reducir el espacio de búsqueda antes de realizar la búsqueda basándose en la comparación entre los n-gramas. En [7], los autores propone la reducción del espacio de búsqueda en la base de la distancia de Kullback-Leibler.

Hay que destacar la importancia de proteger los datos confidenciales de una empresa. Defenderse de intrusos dentro del sistema es una de las prioridades de los encargados de los departamentos de informática de las empresas. Actualmente, existen estudios que proponen utilizar los métodos de detección automática de plagio para la detección de intrusos en el sistema de red [120, 119, 96]. La idea es buscar patrones en los datos de la carga útil tratando los datos como secuencias de bytes. No obstante, es una idea muy reciente que necesita madurar.

Otra línea de investigación muy reciente es la detección de plagio de opiniones[1]. El plagio de opiniones se produce muy frecuentemente en la Web, puesto que la influencia de un blog se mide entre otras cosas a partir del número de opiniones vertidas en él. Incrementar el número de opiniones plagiando opiniones de otros blogs, puede aumentar la influencia de un blog, con lo que se obtendría un beneficio publicitario. Actualmente, las empresas buscan métodos automáticos de detección de plagio de opiniones.

6.2. Cómo beneficiarse de las ventajas de la Web 2.0

Como ya hemos comentado, la Web 2.0 nos aporta un flujo entre empresa y consumidores, el cual aporta un mayor contacto entre la entidad y el cliente. Pero más importante aún, la

[1] http://kdd.di.unito.it/DyNak2010/

Web 2.0 aporta un flujo entre consumidores. Este último flujo puede ser el que más datos aporte a una empresa, puesto que si una empresa sabe como tratar dichos datos tendrá información de primera mano sobre las tendencias de los consumidores.

Sin embargo, la Web 2.0 se ha convertido en una inmensa red de información la cual es imposible de analizar todos los datos que aparecen en ella. Por eso es conveniente que empresas compartan dicha información para obtener un beneficio mutuo. En este trabajo hemos propuesto un algoritmo capaz de analizar las opiniones de los consumidores realizadas en las redes sociales y, compartir y/o intercambiar este análisis a partir de una fusión de ontologías.

Sin embargo, se ha demostrado la dificultad que entraña la tarea de analizar las opiniones en la Web 2.0 de los usuarios de blogs a través de una ontología, principalmente cuando esta ontología esta preestablecida. En estudios anteriores como [130] y [131], las ontologías en cambio se generaban a partir del corpus con lo que facilitaba la búsqueda de conceptos. Sin embargo, creemos que nuestros experimentos están más próximos al problema del mundo real, puesto que las empresas ya poseen de antemano una ontología del dominio. No obstante, hemos comprobado como al realizar el proceso de fusión de ontologías no se pierde prácticamente ningún dato de los anteriormente calculados por el análisis de opiniones.

En cuanto a la fusión de ontologías, hemos comprobado que en el dominio del turismo, con pequeñas ontologías los métodos directos son preferibles frente a los métodos via *upper ontology*. Sin embargo, conforme aumenta el tamaño de las ontologías el proceso se invierte mejorando los resultados con los métodos *upper ontology*. Por otro lado, hemos comprobado que los métodos no estructurados de las técnicas via *uper ontology* obtienen mejores resultados en precisión y F-measure. Finalmente, con respecto a los resultados obtenidos con las *upper ontology*, hemos llegado a la conclusión que utilizando OpenCyc se obtienen mejores resultados en el dominio del turismo que con SUMO.

Por último, hemos comprobado que en las opiniones negativas, los consumidores suelen introducir frases irónicas, provocando que la detención de la polaridad de estos textos sea más dificultosa. Un aspecto interesante a tener en cuenta en futuros trabajos sería introducir algún método automático para detectar la ironía y el sarcasmo. Esto nos ayudaría a poder clasificar correctamente la polaridad, ya que si utilizamos el sarcasmo podemos invertir la polaridad de sus palabras. Detectar automáticamente el sarcasmo sería efectivo sobre todo para mejorar el porcentaje de aciertos en las opiniones negativas.

Bibliografía

[1] Profesionalismo médico en el nuevo milenio. Un estatuto para el ejercicio de la medicina, Federación Europea de Medicina Interna. American College of Physicians y American Board of Internal Medicine. *Revista Medica de Chile*, 131:457–460, 2003. Suscrito por el Colegio Médico de Chile A.G en 2004.

[2] iparadigms: Digital solutions for a new era in information. 2004. http://www.iparadigms.com.

[3] A. Abbasi, H. Chen, and A. Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems*, 26(3):1–34, 2008.

[4] Z. Aleksovski, M. C. A. Klein, W. ten Kate, and F. van Harmelen. Matching unstructured vocabularies using a background ontology. *Proceedings of the 15th International Conference on Managing Knowledge in a World of Networks, (EKAW '06)*, pages 182–197, 2006.

[5] A. Barrón-Cedeño. *Detección automática de plagio en texto.* Tesis de Máster, Universidad Politécnica de Valencia, 2008.

[6] A. Barrón-Cedeño and P. Rosso. On automatic plagiarism detection based on n-grams comparisons. *Proceedings European Conference on Information Retrieval, (ECIR'09)*, pages 696–700, 2009.

[7] A. Barrón-Cedeño, P. Rosso, and J.M. Benedí. Reducing the plagiarism detection search space on the basis of the kullback-leibler distance. *Proceedings of the 10th International Conference on Computational Linguistics and Intelligent Text Processing, (CICLing'09)*, pages 523–534, 2009.

[8] A. Barrón-Cedeño, P. Rosso, D. Pinto, and A. Juan. On cross-lingual plagiarism analysis using a statistical model. *Benno Stein, Efstathios Stamatatos, and Moshe Koppel, editors, ECAI 2008 Workshop on Uncovering Plagiarism, Authorship, and Social Software Misuse (PAN 08)*, pages 9–13, Julio 2008.

[9] N. Casellas, M. Blázquez, A. Kiryakov, P. Casanovas, M. Poblet, and V.R. Benjamins. OPJK into PROTON: Legal domain ontology integration into an upper-level ontology. *Proceedings of OTM Workshops 2005, LNCS 3762*, pages 846–855, 2005.

[10] S. Castano, V. De Antonellis, and S. De Capitani di Vimercati. Global viewing of heterogeneous data sources. *Proceedings of IEEE Transactions on Knowledge and Data Engineering*, 13(2):277–297, 2001.

[11] C. Chen, J. Yeh, and H. Ke. Plagiarism detection using rouge and wordnet. *Computing Research Repositoy (CoRR)*, Marzo 2010.

[12] K.W. Church and P. Hanks. Word association norms, mutual information and lexicography. *Proceedings of the 27th Annual Conference of the Association of Computational Linguistics*, 1989.

[13] S.A. Martínez De La Cruz. Importancia de los sistemas de información para las pequeñas empresas, 2005.

[14] E. Dale and J. Chall. *A formula for predicting readability.* Educ. Res. Bull., 27, 1948.

[15] L. Dey and S. M. Haque. Opinion mining from noisy text data. *Proceedings of the Workshop on Analytics for Noisy Unstructured Text Data, SIGIR 08*, 2008.

[16] J. Dierderich. Computational methods to detect plagiarism in assessment. *Proceedings of the 7th International Conference on Information Technology Based Higher Education and Training (ITHET '06)*, pages 147–154, 2006.

[17] H. Do and E. Rahm. COMA: A system for flexible combination of schema matching approaches. *VLDB '02: Proceedings of the 28th international conference on Very Large Data Bases*, pages 610–621, 2002.

[18] A. Doan, P. Domingos, and A. Halevy. Reconciling schemas of disparate data sources: A machine-learning approach. *Proceedings of Special Interest Group on Management Of Data (SIGMOD'01)*, 2001.

[19] A. Doan, J. Madhavan, R. Dhamankar, P. Domingos, and A. Halevy. Learning to match ontologies on the Semantic Web. *VLDB '03: Proceedings of the Very Large Data Bases Journal*, 12(4):303–319, 2003.

[20] A. Doan, J. Madhavan, P. Domingos, and A. Halevy. Ontology matching: a machine learning approach. *Handbook of ontologies, International handbooks on information systems*, pages 385–404, 2004.

[21] Z. Dong and Q. Dong. *Hownet And the Computation of Meaning.* World Scientific Publishing Co., Inc., River Edge, NJ, 2006.

[22] H. Dreher. Automatic conceptual analysis for plagiarism detection. *Journal of Issues in Informing Science and Information Technology 4*, pages 601–614, 2007.

[23] M. Ehrig and S. Staab. QOM - Quick Ontology Matching. *Proceedings of International Semantic Web Conference (ISWC)*, pages 683–697, 2004.

[24] A. Esuli and F. Sebastiani. SENTIWORDNET: A publicly available lexical resource for opinion mining. *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC06)*, pages 417–422, 2006.

[25] J. Euzenat. An api for ontology alignment. *Proceedings of 3rd Conference on International Semantic Web Conference (ISWC)*, pages 698–712, 2004.

[26] J. Euzenat, P. Guégan, and P. Valtchev. OLA in the OAEI 2005 alignment contest. *Proceedings of the K-CAP Workshop on Integrating Ontologies*, pages 61–71, 2005.

[27] J. Euzenat and P. Shvaiko. *Ontology matching*. Springer, 2007.

[28] C. Fellbaum. *WordNet: An Electronic Lexical Database*. MIT Press, 1998.

[29] R. Flesch. A new readability yardstick. *Journal of Applied Psychology*, (32):221–233, 1948.

[30] A. Flint, S. Clegg, and R. Macdonald. Exploring staff perceptions of student plagiarism. *Journal of Further and Higher Education*, 30(2):145–156, 2006.

[31] M. Gamon, A. Aue, S. Corston-Oliver, and E. Ringger. Pulse: Mining customer opinions from free text. *IDA, Vol 3646 of Lecture Notes in Computer Science*, pages 121–132, 2005.

[32] A. Gangemi, N. Guarino, C. Masolo, A. Oltramari, and L. Schneider. Sweetening ontologies with DOLCE. *Proceedings of Knowledge Engineering and Knowledge Management by the Masses (EKAW'02)*, pages 166–181, 2002.

[33] V. Gil and F. Romero. *Crossumer, Claves para entender al consumidor español de nueva generación*. Ediciones Gestión 2000, 2008.

[34] S. J. Girón Castro. *Anotaciones sobre el plagio*. Universidad Sergio Arboleda, 2008.

[35] F. Giunchiglia, P. Shvaiko, and M. Yatskevich. S-Match: an algorithm and an implementation of semantic matching. *Proceedings of European Semantic Web Symposium (ESWS'04)*, pages 61–75, 2004.

[36] P. Grenon, B. Smith, and L. Goldberg. Biodynamic ontology: applying BFO in the biomedical domain. *Ontologies in Medicine: Studies in Health Technology and Informatics*, pages 20–38, 2004.

[37] C. Grozea, C. Gehl, and M. Popescu. Encoplot: Pairwise sequence matching in linear time applied to plagiarism detection. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 10–18, 2009.

[38] L. Harris and A. Rae. Social networks: The future of marketing for small business. *The Journal of Business Strategy*, 2009.

[39] V. Hatzivassiloglou and K. R. McKeown. Predicting the semantic orientation of adjectives. *Proceedings of the Association for Computational Linguistics (ACL'97)*, pages 174–181, 1997.

[40] V. Hatzivassiloglou and J. Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'00)*, 2000.

[41] H. Herre, B. Heller, P. Burek, R. Hoehndorf, F. Loebe, and H. Michalek. General formal ontology (GFO): A foundational ontology integrating objects and processes. Part I: Basic principles. Technical Report Nr. 8, Research Group Ontologies in Medicine (Onto-Med), Univ. Leipzig, 2006.

[42] D. Hindle. Noun classification from predicate argument structures. *Proceedings of the 28th Annual Meeting of the ACL*, pages 268–275, 1990.

[43] D. I. Holmes. A stylometric analysis of mormon scripture and related texts. *Journal of the Royal Statistical Society Series A Statistics in Society*, 155(1):91–120, 1992.

[44] A. Honore. Some simple measures of richness of vocabulary. *Association for Literary and Linguistic Computing Bulletin*, 7(2):172–177, 1979.

[45] G. Hooley, G. Greenley, and V. Wong. Marketing: A history of the next decade. *Journal of Marketing Management*, 2003.

[46] M. Hu and B. Liu. Mining and summarizing customer reviews. *KDD '04: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 168–177, 2004.

[47] R.A. Hummel and S.W. Zucker. On the foundations of relaxation labeling processes. *Proceedings of IEEE Transactions on Pattern Analysis and Machine Intelligence (PAMI)*, 1983.

[48] R. Irribarne and H. Retondo. Plagio de obras literarias. ilícitos civiles y penales en derecho de autor. *Instituto Interamericano de Derecho de Autor (Interamerican Copyright Institute - IIDA)*, 1981.

[49] Chen J. and J-Y Nie. Parallel web text mining for cross-language. *IR. Algorithmica*, 28(2):217–241, 2000.

[50] B. Jansen, M. Zhang, K. Sobel, and A. Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 2009.

[51] J. Kamps and M. Marx. Words with attitude. *Proceedings of the 1st International WordNet Conference*, pages 332–341, 2002.

[52] N. Kang, A. Gelbukh, and S. Y. Han. Ppchecker: Plagiarism pattern checker in document copy detection. *Lecture notes in computer science*, (4188):661–668, 2006.

[53] P. Kim. The Forrester Wave: Brand monitoring, Q3 2006, 2006. Forrester Wave (white paper).

[54] S. Kim and E. Hovy. Determining the sentiment of opinions. *Proceedings of the 18th International Conference on Computational Linguistics (COLING'04)*, pages 1267–1373, 2004.

[55] J. Kincaid, R. Fishburne, R. Rogers, and B. Chissom. Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel. *Research Branch Report 8Ú75 Millington TN: Naval Technical Training US Naval Air Station*, 1975.

[56] A. Kiryakov, K.I. Simov, and M. Dimitrov. OntoMap: portal for upper-level ontologies. *Proceedings of the International Conference on Formal Ontology in Information Systems (FOIS'01)*, pages 47–58, 2001.

[57] L.A. Kloda and K.Ñicholson. Plagiarism detection software and academic integrity: The canadian perspective. *Proceedings Librarians Information Literacy Annual Conference (LILAC)*, 2005.

[58] K. Kotis, G.A. Vouros, and K. Stergiou. Capturing semantics towards automatic coordination of domain ontologies. *Proceedings of the 11th International conference of Artificial Intelligence: Methodology, Systems, Architectures - Semantic Web Challenges - AIMSA 2004*, pages 22–32, 2004.

[59] P. Lambrix and H. Tan. SAMBO-A system for aligning and merging biomedical ontologies. *Web Semantics: Science, Services and Agents on the World Wide Web*, 4(3):196–206, 2006.

[60] H.D. Lasswell and J.Z. Namenwirth. *The Lasswell Value Dictionary*. New Haven: Yale University Press, 1969.

[61] B.T. Le, R. Dieng-Kuntz, and F. Gandom. On ontology matching problems - for building a corporate semantic web in a multi-communities organization. *Proceedings of the Sixth International Conference on Enterprise Information Systems (ICEIS'04)*, (4):236–243, Abril 2004.

[62] D. Lee, O. R. Jeong, and S. G. Lee. Opinion mining of customer feedback data on the web. *Proceedings of the 2nd international conference on Ubiquitous Information Management and Communication*, 2008.

[63] D. H. Lehmer. Arithmetical periodicities of bessel functions. *Annals of Mathematics*, 33:143–150, 1932.

[64] D.B. Lenat and R.V. Guha. *Building Large Knowledge-Based Systems; Representation and Inference in the Cyc Project*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1989.

[65] V.I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966.

[66] J. Li. LOM: A lexicon-based ontology mapping tool. *Proceedings of the Workshop on Performance Metrics for Intelligent Systems, (PerMIS'04)*, 2004.

[67] B. Liu, M. Hu, and J. Cheng. Opinion Observer: Analysing and comparaing opinions on the Web. *Proceedings of International World Wide Web Conference (WWW05)*, 2005.

[68] A. Locoro. Ontology Matching using Upper Ontologies and Natural Language Processing. In *PhD-Thesis Course in Electronic and Computer Engineering, Robotics and Telecommunications*, 2010. Università di Genova, Italia.

[69] C. Lyon, R. Barrett, and J. Malcolm. A theoretical basis to the automated detection of copying between texts, and its practical implementation in the ferret plagiarism and collusion detector. 2004.

[70] C. Lyon, R. Barrett, and J. Malcolm. Plagiarism is easy, but also easy to detect. *Plagiary: Cross-Disciplinary Studies in Plagiarism, Fabrication, and Falsification*, 1, 2006.

[71] J. Madhavan, P. A. Bernstein, and E. Rahm. Generic schema matching with Cupid. *Proceedings of the 27th International Conference on Very Large Data Bases (VLDB'01)*, pages 49–58, 2001.

[72] B. Magnini, Luciano Serafini, and M. Speranza. Making explicit the semantics hidden in schema models. *Proceedings of ISWC workshop on Human Language Technology for the Semantic Web and Web Services*, 2003.

[73] P.C. Mahalanobis. On the generalised distance in statistics. *Proceedings of the National Institute of Science of India*, pages 49–55, 1936.

[74] J. Malcolm and P.C.R. Lane. Efficient search for plagiarism on the web. *Proceedings of The International Conference on Technology, Communication and Education*, pages 206–211, 2008.

[75] J. Malcolm and P.C.R Lane. Tackling the pan'09 external plagiarism detection corpus with a desktop plagiarism detector. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 29–33, 2009.

[76] V. Mascardi, V. Cordì, and P. Rosso. A comparison of upper ontologies. *Atti del Workshop Dagli Oggentti agli Agenti, WOA*, pages 55–64, 2007.

[77] V. Mascardi, A. Locoro, and P. Rosso. Automatic ontology matching via upper ontologies: A systematic evaluation. *IEEE Transactions on Knowledge and Data Engineering*, 99(1), 2009. doi: 10.1109/TKDE.2009.154.

[78] V. Mascardi, P. Rosso, and V. Cordì. Enhancing communication inside multiagent systems - an approach based on alignment via upper ontologies. *Proceedings of the Multi-Agent Logics, Languages, and Organisations - Federated Workshops, (MALLOW-AWESOME'007)*, pages 92–107, 2007.

[79] V. H. Medina García and L. Sánchez Gracía. Efectos negativos de la web en la ética y la sociedad. Technical report, Universidad Pontificia de Salamanca.

[80] S. Melnik, H. Garcia-Molina, and E. Rahm. Similarity Flooding: A versatile graph matching algorithm and its application to schema matching. *Proceedings of the 18th International Conference on Data Engineering (ICDE'02)*, page 117, 2002.

[81] J. Mira. Tu opinión vale (mucho) dinero. El Periódico, Enero 2009.

[82] A. Monge and C. Elkan. The field-matching problem: algorithm and applications. *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, 1996.

[83] M. Muhr, M. Zechner, R. Kern, and M. Granitzer. External and intrinsic plagiarism detection using vector space models. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 47–55, 2009.

[84] T.Ñasukawa and J. Yi. Sentiment analysis: capturing favorability using natural language processing. *Proceedings of the International Conference on Knowledge Capture (K-CAP'03)*, pages 70–77, 2003.

[85] A.Ñiculcea, J. Whelan, J. Slama, M. Cancho-Rosado, B. Díaz-Palomo, C. Rodríguez-Agudín, and V. Sánchez-Muñoz. Web 2.0: El negocio de las redes sociales. Technical report, 2007.

[86] I. Niles and A. Pease. Towards a standard upper ontology. *Proceedings of the international conference on Formal Ontology in Information Systems (FOIS'01)*, pages 2–9, 2001.

[87] N.F. Noy and M.A. Musen. The PROMPT suite: interactive tools for ontology merging and mapping. *International Journal of Human-Computer Studies*, 59(6):983–1024, 2003.

[88] C. E. Osgood, G. J. Suci, and P. H. Tannenbaum. *The Measurement of Meaning*. University of Illinois Press, Chicago, 1957.

[89] R. Philip. Mining the web for bilingual text. *Proceedings of the Association for Computational Linguistics (ACL'99)*, 1999.

[90] C. Phytila. An Analysis of the SUMO and Description in Unified Modeling Language. 2002. no publicado.

[91] A. Popescu and O. Etzioni. Extracting product features and opinions from reviews. *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05)*, pages 339–346, 2005.

[92] M. Potthast, B. Stein, and M. Anderka. A Wikipedia-based multilingual retrieval model. *MacDonald, Ounis, Plachouras, Ruthven and White (eds.). 30th European Conference Recent Advances in Natural Language Processing (RALNP'03)*, pages 401–408, 2003.

[93] M. Potthast, B. Stein, A. Eiselt, A. Barrón-Cedeño, and P. Rosso. Overview of the 1st International Competition on Plagiarism Detection. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 1–9, 2003.

[94] B. Pouliquen, R. Steinberger, and C. Ignat. Automatic identification of document translations in large multilingual document collections. *Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP'03)*, pages 401–408, 2003.

[95] A. Reyes, P. Rosso, and D. Buscaldi. Humor in the blogosphere: First clues for a verbal humor taxonomy. *Journal of Intelligent Systems*, 18(4), 2009.

[96] K. Rieck and P. Laskov. Linear-time computation of similarity measures for sequential data. *The Journal of Machine Learning Research*, 9:23–48, 2008.

[97] E. Rosselot, M. Bravo, M. Kottow, C. Valenzuela, M. O'Ryan, S. Thambi, and et al. En referencia al plagio intelectual. documento de la comisión de Ética de la facultad de medicina de la universidad de chile. *Revista Médica de Chile*, 136(5):653–658, 2008.

[98] F. Uribe Saavedra. Las redes sociales digitales, una herramienta de marketing para las pymes. Technical report, 2010. Revisión de literatura.

[99] Bryan Scaife. Evaluation of plagiarism detection software. Technical report, IT Consultancy, 2007.

[100] S.K. Semy, M.K. Pulvermacher, and L.J. Obrst. Toward the use of an upper ontology for U.S. government and U.S. military domains: An evaluation. *Submission to Workshop on IIWeb*, 2004.

[101] N. Shivakumar and H. Garcia-Molina. Scam: A copy detection mechanism for digital documents. *Proceedings of the 2nd International Conference on Theory and Practice of Digital Libraries*, 1995.

[102] H. S. Sichel. On a Distribution Law for Word Frequencies. *Journal of the American Statistical Association*, 70(351):542–547, 1975.

[103] D. Smallbone, R. Leig, and D.Ñorth. The characteristics and strategies of high growth smes. *International Journal of Entrepreneurial Behaviour Research*, 1995.

[104] J.F. Sowa. *Knowledge Representation: Logical, Philosophical, and Computational Foundations*. Brooks Cole Publishing, 2000.

[105] E. Stamatatos. Intrinsic plagiarism detection using character n-gram profiles. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 38–46, 2009.

[106] B. Stein and S. Meyer Zu Eissen. Topic identification: Framework and application. *Proceedings of the 4th International Conference on Knowledge Management, Journal of Universal Computer Science, Know-Center (I-KNOW'04)*, pages 353–360, 2004.

[107] B. Stein and S. Meyer Zu Eissen. Intrinsic plagiarism detection. *Mounia Lalmas, Andy MacFarlane, Stefan M. Rüger, Anastasios Tombros, Theodora Tsikrika, and Alexei Yavlinsky, editors, ECIR, volume 3936 of Lecture Notes in Computer Science*, pages 565–569, 2006.

[108] B. Stein and S. Meyer Zu Eissen. Intrinsic plagiarism analysis with meta learning. *Proceedings of the SIGIR'07 Workshop on Plagiarism Analysis, Authorship Identification, and Near-Duplicate Detection (PAN 07)*, pages 45–50, 2007.

[109] B. Stein, S. Meyer zu Eissen, and M. Kulig. Plagiarism detection without reference collections. *R. Decker and H. Lenz, editors, Advances in Data Analysis*, pages 359–366, 2007.

[110] G. Stoilos, G.B. Stamou, and S. Kollias. A string metric for ontology alignment. *Proceedings of the International Semantic Web Conference (ISWC'05)*, pages 624–637, 2005.

[111] P. J. Stone, D. C. Dunphy, Smith M. S., and D. M. Ogilvie. *The General Inquirer: A Computer Approach to Content Analysis*. MIT Press, Cambridge, MA, 1966.

[112] D. Storey. *Understanding the Small Business Sector*. Routledge, 1994.

[113] C. Strapparava and A. Valitutti. WordNet-Affect: an affective extension of WordNet. *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC'04)*, 4:1083–1086, 2004.

[114] W. Sutherland-Smith and R. Carr. Turnitin.com: Teachers' perspectives of anti-plagiarism software in raising issues of educational integrity. *Journal of University Teaching and Learning Practice*, 2(3), 2005.

[115] P. D. Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL'02)*, pages 417–424, 2002.

[116] P. D. Turney and M. L. Littman. Measuring praise and criticism: Inference of semantic orientation from association. *Proceedings of the Transactions On Information Systems (TOIS'03)*, 21(4):315–346, 2003.

[117] A. Utsumi. A unified theory of irony and its computational formalization. *Proceedings of the 16th conference on Computational linguistics*, pages 962–967, 1996.

[118] E. Valles Balaguer. Putting ourselves in sme's shoes: Automatic detection of plagiarism by the wcopyfind tool. *Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software*, pages 34–35, 2009.

[119] K. Wang, J. J. Parekh, and S. J. Stolfo. Anomalous payload-based network intrusion detection. *Proceedings of the 9 th International Symposium on Recent Advances in Intrusion Detection (RAID'06)*, pages 226–248, 2006.

[120] K. Wang and S.J. Stolfo. Anomalous payload-based network intrusion detection. *International symposium on recent advances in intrusion detection Nº 7*, 3224:203–222, 2004.

[121] J. Wiebe. Learning subjective adjectives from corpora. *Proceedings of the Innovative Applications of Artificial Intelligence Conference (AAAI/IAAI)*, pages 735–740, 2000.

[122] J. Wiebe and E. Riloff. Creating subjective and objective sentence classifiers from unannotated texts. *Proceedings of International Conference on Intelligent Text Processing and Computational Linguistics*, 2005.

[123] T. Wilson, J. Wiebe, and R. Hwa. Just how mad are you? finding strong and weak opinion clauses. *Proceedings of the 19th national conference on Artifical Intelligence (AAAI'04)*, pages 761–769, 2004.

[124] W. E. Winkler. The state of record linkage and current research problems. *Statistics of Income Division, Internal Revenue Service Publication R99/04.*, 1999.

[125] L. Yu, J. Ma, S. Tsuchiya, and F. Ren. Opinion mining: A study on semantic orientation analysis for online document. *Proceedings of the 7th World Congress on Intelligent Control and Automation*, June 2008.

[126] G. Yule. *The Statistical Study of Literary Vocabulary.* Cambridge University Press, 1944.

[127] J. Zabin and A. Jefferies. Social media monitoring and analysis: Generating consumer insights from online conversation. Aberdeen Group Benchmark Report, January 2008.

[128] C. Zdenek, T. Michal, and J. Karel. Multilingual plagiarism detection. *AIMSA'08: Proceedings of the 13th international conference on Artificial Intelligence*, pages 83–92, 2008.

[129] W. Zhang and S. Watts. Online communities as communities of practice: A case study. *Journal of Knowledge Management*, 2008.

[130] L. Zhao and C. Li. Ontology based opinion mining for movie reviews. *Proceedings of the International Conference on Knowledge Science, Engineering and Management (KSEM'09)*, pages 204–214, 2009.

[131] L. Zhou and P. Chaovalit. Ontology-supported polarity mining. *Journal of the American Society for Information Science and Technology*, 59(1):98–110, 2008.

[132] G. K. Zipf. *Human Behavior and the Principle of Least-Effort*. Addison-Wesley, 1949.

Apéndice A

Resultados de los experimentos de fusión de ontologías

Distancia	**Upper Onto**	**Método**	**Encontrados**	**Correc**	**Prec**	**Rec**	**F-Meas**
Manual	None	None	16	16	1.00	1.00	1.00
equal	None	None	16	6	0.38	0.38	0.38
SMOA	None	None	43	10	0.23	**0.63**	0.34
Levenshtein	None	None	39	10	0.26	**0.63**	0.36
equal	SUMO	NoStruct	2	2	**1.00**	0.13	0.22
SMOA	SUMO	NoStruct	83	10	0.12	**0.63**	0.20
Levenshtein	SUMO	NoStruct	14	5	0.36	0.31	0.33
equal	SUMO	Struct	3	2	0.67	0.13	0.21
SMOA	SUMO	Struct	94	10	0.11	**0.63**	0.18
Levenshtein	SUMO	Struct	17	5	0.29	0.31	0.30
equal	OpenCyc	NoStruct	3	2	0.67	0.13	0.21
SMOA	OpenCyc	NoStruct	54	9	0.17	0.56	0.26
Levenshtein	OpenCyc	NoStruct	12	6	0.50	0.38	**0.43**
equal	OpenCyc	Struct	10	2	0.20	0.13	0.15
SMOA	OpenCyc	Struct	78	9	0.12	0.56	0.19
Levenshtein	OpenCyc	Struct	29	6	0.21	0.38	0.27

Tabla A.1: Fusión de ETP-Tourism y L_Tour

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	35	35	1.00	1.00	1.00
equal	None	None	18	15	**0.83**	0.43	**0.57**
SMOA	None	None	29	17	0.59	**0.49**	0.53
Levenshtein	None	None	18	15	**0.83**	0.43	**0.57**
equal	SUMO	NoStruct	9	7	0.78	0.20	0.32
SMOA	SUMO	NoStruct	89	16	0.18	0.46	0.26
Levenshtein	SUMO	NoStruct	37	15	0.41	0.43	0.42
equal	SUMO	Struct	14	7	0.50	0.20	0.29
SMOA	SUMO	Struct	133	16	0.12	0.46	0.20
Levenshtein	SUMO	Struct	52	15	0.29	0.43	0.34
equal	OpenCyc	NoStruct	10	6	0.60	0.17	0.27
SMOA	OpenCyc	NoStruct	37	16	0.43	0.46	0.44
Levenshtein	OpenCyc	NoStruct	26	17	0.65	**0.49**	0.56
equal	OpenCyc	Struct	27	6	0.22	0.17	0.19
SMOA	OpenCyc	Struct	145	16	0.11	0.46	0.18
Levenshtein	OpenCyc	Struct	59	17	0.29	**0.49**	0.36

Tabla A.2: Fusión de ETP-Tourism y Tourism-ProtegeExportOWL

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	98	98	1.00	1.00	1.00
equal	None	None	195	80	0.41	0.82	0.55
SMOA	None	None	221	84	0.38	**0.86**	0.52
Levenshtein	None	None	205	82	0.40	0.84	0.54
equal	SUMO	NoStruct	18	14	**0.78**	0.14	0.24
SMOA	SUMO	NoStruct	415	83	0.20	0.85	0.32
Levenshtein	SUMO	NoStruct	264	74	0.28	0.76	0.41
equal	SUMO	Struct	20	14	0.70	0.14	0.24
SMOA	SUMO	Struct	461	83	0.18	0.85	0.30
Levenshtein	SUMO	Struct	264	74	0.28	0.76	0.41
equal	OpenCyc	NoStruct	38	16	0.42	0.16	0.24
SMOA	OpenCyc	NoStruct	143	80	0.56	0.82	0.67
Levenshtein	OpenCyc	NoStruct	122	78	0.64	0.80	**0.71**
equal	OpenCyc	Struct	53	16	0.30	0.16	0.21
SMOA	OpenCyc	Struct	200	80	0.40	0.82	0.54
Levenshtein	OpenCyc	Struct	144	78	0.54	0.80	0.64

Tabla A.3: Fusión de ETP-Tourism y qallme-tourism

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	26	26	1.00	1.00	1.00
equal	None	None	15	12	0.80	0.46	0.59
SMOA	None	None	17	13	0.76	0.50	**0.60**
Levenshtein	None	None	17	13	0.76	0.50	**0.60**
equal	SUMO	NoStruct	7	7	**1.00**	0.27	0.42
SMOA	SUMO	NoStruct	34	13	0.38	0.50	0.43
Levenshtein	SUMO	NoStruct	21	13	0.62	0.50	0.55
equal	SUMO	Struct	13	7	0.54	0.27	0.36
SMOA	SUMO	Struct	62	13	0.21	0.50	0.30
Levenshtein	SUMO	Struct	30	13	0.43	0.50	0.46
equal	OpenCyc	NoStruct	7	5	0.71	0.19	0.30
SMOA	OpenCyc	NoStruct	26	14	0.54	**0.54**	0.54
Levenshtein	OpenCyc	NoStruct	15	12	0.80	0.46	0.59
equal	OpenCyc	Struct	17	5	0.29	0.19	0.23
SMOA	OpenCyc	Struct	70	14	0.20	**0.54**	0.29
Levenshtein	OpenCyc	Struct	36	12	0.33	0.46	0.39

Tabla A.4: Fusión de ETP-Tourism y TravelOntology

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	4	4	1.00	1.00	1.00
equal	None	None	1	1	**1.00**	0.25	0.40
SMOA	None	None	2	1	0.50	0.25	0.33
Levenshtein	None	None	1	1	**1.00**	0.25	0.40
equal	SUMO	NoStruct	1	1	**1.00**	0.25	0.40
SMOA	SUMO	NoStruct	13	2	0.15	**0.50**	0.24
Levenshtein	SUMO	NoStruct	1	1	**1.00**	0.25	0.40
equal	SUMO	Struct	2	1	0.50	0.25	0.33
SMOA	SUMO	Struct	17	2	0.12	**0.50**	0.19
Levenshtein	SUMO	Struct	2	1	0.50	0.25	0.33
equal	OpenCyc	NoStruct	0	0	0.00	0.00	0.00
SMOA	OpenCyc	NoStruct	5	2	0.40	**0.50**	**0.44**
Levenshtein	OpenCyc	NoStruct	1	1	**1.00**	0.25	0.40
equal	OpenCyc	Struct	0	0	0.00	0.00	0.00
SMOA	OpenCyc	Struct	18	2	0.11	**0.50**	0.17
Levenshtein	OpenCyc	Struct	7	1	0.14	0.25	0.18

Tabla A.5: Fusión de Tourism-ProtegeExportOWL y L_Tour

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	18	18	1.00	1.00	1.00
equal	None	None	18	7	0.39	0.39	**0.39**
SMOA	None	None	80	12	0.15	**0.67**	0.24
Levenshtein	None	None	43	10	0.23	0.56	0.32
equal	SUMO	NoStruct	2	2	**1.00**	0.11	0.20
SMOA	SUMO	NoStruct	92	12	0.13	**0.67**	0.22
Levenshtein	SUMO	NoStruct	9	4	0.44	0.22	0.30
equal	SUMO	Struct	2	2	**1.00**	0.11	0.20
SMOA	SUMO	Struct	92	12	0.13	**0.67**	0.22
Levenshtein	SUMO	Struct	10	4	0.40	0.22	0.29
equal	OpenCyc	NoStruct	3	2	0.67	0.11	0.19
SMOA	OpenCyc	NoStruct	92	11	0.12	0.61	0.20
Levenshtein	OpenCyc	NoStruct	2	6	0.38	0.33	0.35
equal	OpenCyc	Struct	6	2	0.33	0.11	0.17
SMOA	OpenCyc	Struct	100	11	0.11	0.61	0.18
Levenshtein	OpenCyc	Struct	20	6	0.30	0.33	0.32

Tabla A.6: Fusión de qallme-tourism y L_Tour

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	26	26	1.00	1.00	1.00
equal	None	None	13	11	0.85	**0.42**	**0.56**
SMOA	None	None	13	11	0.85	**0.42**	**0.56**
Levenshtein	None	None	13	11	0.85	**0.42**	**0.56**
equal	SUMO	NoStruct	7	6	**0.86**	0.23	0.36
SMOA	SUMO	NoStruct	39	11	0.28	**0.42**	0.34
Levenshtein	SUMO	NoStruct	29	10	0.34	0.38	0.36
equal	SUMO	Struct	10	6	0.60	0.23	0.33
SMOA	SUMO	Struct	55	11	0.20	**0.42**	0.27
Levenshtein	SUMO	Struct	42	10	0.24	0.38	0.30
equal	OpenCyc	NoStruct	6	3	0.50	0.12	0.19
SMOA	OpenCyc	NoStruct	27	11	0.41	**0.42**	0.42
Levenshtein	OpenCyc	NoStruct	15	11	0.73	**0.42**	0.54
equal	OpenCyc	Struct	16	3	0.19	0.12	0.14
SMOA	OpenCyc	Struct	100	11	0.11	**0.42**	0.18
Levenshtein	OpenCyc	Struct	35	11	0.31	**0.42**	0.36

Tabla A.7: Fusión de qallme-tourism y Tourism-ProtegeExportOWL

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	22	22	1.00	1.00	1.00
equal	None	None	12	7	0.58	0.32	0.41
SMOA	None	None	15	8	0.53	0.36	**0.43**
Levenshtein	None	None	15	8	0.53	0.36	**0.43**
equal	SUMO	NoStruct	5	5	**1.00**	0.23	0.37
SMOA	SUMO	NoStruct	16	8	0.50	0.36	0.42
Levenshtein	SUMO	NoStruct	17	8	0.47	0.36	0.41
equal	SUMO	Struct	9	5	0.56	0.23	0.32
SMOA	SUMO	Struct	33	8	0.24	0.36	0.29
Levenshtein	SUMO	Struct	24	8	0.33	0.36	0.35
equal	OpenCyc	NoStruct	4	1	0.25	0.05	0.08
SMOA	OpenCyc	NoStruct	16	8	0.50	0.36	0.42
Levenshtein	OpenCyc	NoStruct	13	7	0.54	0.32	0.40
equal	OpenCyc	Struct	9	1	0.11	0.05	0.06
SMOA	OpenCyc	Struct	37	9	0.24	**0.41**	0.30
Levenshtein	OpenCyc	Struct	25	7	0.28	0.32	0.30

Tabla A.8: Fusión de qallme-tourism y TravelOntology

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	4	4	1.00	1.00	1.00
equal	None	None	2	2	**1.00**	0.50	**0.67**
SMOA	None	None	7	3	0.43	**0.75**	0.55
Levenshtein	None	None	3	2	0.67	0.50	0.57
equal	SUMO	NoStruct	0	0	0.00	0.00	0.00
SMOA	SUMO	NoStruct	5	2	0.40	0.50	0.44
Levenshtein	SUMO	NoStruct	2	2	**1.00**	0.50	**0.67**
equal	SUMO	Struct	0	0	0.00	0.00	0.00
SMOA	SUMO	Struct	9	2	0.22	0.50	0.31
Levenshtein	SUMO	Struct	4	2	0.50	0.50	0.50
equal	OpenCyc	NoStruct	1	1	**1.00**	0.25	0.40
SMOA	OpenCyc	NoStruct	9	2	0.22	0.50	0.31
Levenshtein	OpenCyc	NoStruct	2	2	**1.00**	0.50	**0.67**
equal	OpenCyc	Struct	1	1	**1.00**	0.25	0.40
SMOA	OpenCyc	Struct	15	2	0.13	0.50	0.21
Levenshtein	OpenCyc	Struct	2	2	**1.00**	0.50	**0.67**

Tabla A.9: Fusión de TravelOntology y L_Tour

Distancia	Upper Onto	Método	Encontrados	Correc	Prec	Rec	F-Meas
Manual	None	None	17	17	1.00	1.00	1.00
equal	None	None	9	8	**0.89**	0.47	0.62
SMOA	None	None	13	10	0.77	**0.59**	**0.67**
Levenshtein	None	None	11	9	0.82	0.53	0.64
equal	SUMO	NoStruct	9	7	0.78	0.41	0.54
SMOA	SUMO	NoStruct	25	9	0.36	0.53	0.43
Levenshtein	SUMO	NoStruct	15	8	0.53	0.47	0.50
equal	SUMO	Struct	15	7	0.47	0.41	0.44
SMOA	SUMO	Struct	35	9	0.26	0.53	0.35
Levenshtein	SUMO	Struct	20	8	0.40	0.47	0.43
equal	OpenCyc	NoStruct	6	4	0.67	0.24	0.35
SMOA	OpenCyc	NoStruct	17	10	0.59	**0.59**	0.59
Levenshtein	OpenCyc	NoStruct	14	9	0.64	0.53	0.58
equal	OpenCyc	Struct	13	4	0.31	0.24	0.27
SMOA	OpenCyc	Struct	40	10	0.25	**0.59**	0.35
Levenshtein	OpenCyc	Struct	27	9	0.33	0.53	0.41

Tabla A.10: Fusión de TravelOntology y Tourism-ProtegeExportOWL

Apéndice B

Publicaciones en el marco de la investigación

Las investigaciones realizadas en este trabajo ha permitido la publicación de los siguientes artículos:

- Enrique Vallés Balaguer. Putting Ourselves in SME's Shoes: Automatic Detection of Plagiarism by the WCopyFind tool. In Proceedings of the SEPLN'09 Workshop on Uncovering Plagiarism, Authorship and Social Software, CEUR-WS.org, vol. 502, pp. 34-35, 2009, ISSN http://ceur-ws.org/Vol-502, ISSN 1613-0073.
- Enrique Vallés Balaguer, Paolo Rosso, Angela Locoro and Viviana Mascardi. Análisis de opiniones con ontologías. In: POLIBITS, Research journal on Computer science and computer engineering with applications, Num. 41, pp. 29-37, 2010, ISSN 1870-9044.
- Enrique Vallés Balaguer, Paolo Rosso. Empresa 2.0: Detección de plagio y análisis de opiniones. SEPLN'10 Workshop PLN en empresas: Visionando los próximos 10 años (en fase de revisión).