Empresa 2.0: Detección de plagio y análisis de opiniones

Enterprise 2.0: Plagiarism detection and opinion analysis

Enrique Vallés Balaguer
Corex Building Knowledge Solutions
NLE Lab. - ELiRF
Universidad Politécnica de Valencia
evalles@dsic.upv.es

Paolo Rosso*
NLE Lab. - ELiRF
Universidad Politécnica de Valencia
prosso@dsic.upv.es

Resumen: En la era de los medios digitales, las empresas deben afrontar nuevos retos. En este artículo nos ponemos en la piel de una empresa para estudiar cómo afrontar algunos de estos retos como son la protección contra el plagio y el análisis de las opiniones de los consumidores.
Palabras clave: Detección de plagio, análisis de opiniones, fusión de ontologías, Web 2.0, Empresa 2.0

Abstract: In the era of digital media, enterprises face new challenges. In this article, we put ourselves in an enterprise's shoes to study how to resolve some of the challenges such as plagiarism protection and consumer's opinions analysis.
Keywords: Plagiarism detection, opinion analysis, ontology matching, Web 2.0, Enterprise 2.0

1. Introducción

La llegada de la Web 2.0 ha supuesto un nuevo reto para las empresas. Actualmente, aquellas empresas que han apostado por el marketing en los medios digitales, como blogs y redes sociales, son las que mayores posibilidades de éxito tienen en un mercado competitivo, y cada vez más exigente. Sin embargo, la Web 2.0 ha propiciado ciertas conductas entre algunas empresas muy poco recomendables.

Y es que si por una parte, los medios digitales permiten a las empresas tener un mayor contacto con el consumidor informándole de sus productos y servicios; por otra parte, dicha información no sólo está al alcance de la mano de los consumidores, sino que también lo está para las empresas competidoras. Por desgracia, existen empresas que 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 medios digitales a las empresas son mucho mayores que las desventajas. Una de las ventajas está relacionada con las opiniones que comparten los consumidores. Conseguir analizar estas opiniones es de suma 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.

2. Plagio en las empresas

2.1. Prevención de pérdida de datos

El resultado de una pérdida de datos equivale a 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. 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, como (Rieck y Laskov, 2008).

2.2. Plagio de ideas

Para poder acercarse al consumidor, las empresas crean páginas web donde introducen información propia de la empresa, publicitan sus productos y sus servicios. Sin embargo, cuando una empresa lanza una herramienta nueva, introduce una funcionalidad original, tanto consumidores como competidores lo descubren en pocas horas o días.

* El trabajo se engloba dentro del proyecto del MICINN: *TEXT-ENTERPRISE 2.0: Técnicas de Comprensión de textos aplicadas a las necesidades de la Empresa 2.0* (TIN2009-13391-C04-03)

Software	Precision	Recall
Grozea et al.	0,7418	0,6585
WCopyFind	0,0136	0,4586
Ferret	0,0290	0,6048

Tabla 1: Resultados obtenidos en la competición PAN'09

Si una empresa quiere estar en primera línea de salida, debe de estar atenta a sus competidores, para descubrir las novedades y el efecto que tienen en los consumidores, y de esta forma poder mejorar los productos o herramientas que ofrecen. Pero no todas las empresas realizan una competencia leal, sino que existen empresas que utilizan la información que introducen otras en sus páginas web para copiar las ideas de éstas.

2.3. Herramientas para la detección de plagio

Actualmente, hay disponibles herramientas de detección automática de plagio que una empresa puede utilizar para protegerse. Una de estas herramientas es WCopyFind[1]. WCopyFind 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 (Dreher, 2007).

Para comprobar la eficacia de las herramientas para la detección de plagio disponibles en la Web, como WCopyFind, hemos participado en la competición *1st International Competition on Plagiarism Detection*[2] (PAN'09). 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 correspondientes en los documentos originales.

La tabla 1 muestra los resultados que hemos obtenido con el corpus de la competición con la herramienta WCopyFind. También muestra los resultados obtenidos por el equipo que utilizó otra herramienta disponible, Ferret[3].

Observando los resultados, podemos comprobar que para ambas herramientas, los resultados no son buenos comparados con los del ganador de la competición (Potthast et al., 2003). Queremos hacer hincapié en que los resultados de la medida de precisión son muy bajos. Esto es debido principalmente a que las herramientas disponibles no pueden encontrar plagio cuando, por ejemplo hay traducciones a idiomas diferentes al del documento original. Otro factor desfavorable añadido que tiene WCopyFind es que tampoco se tiene en cuenta la modificación de palabras, como pueden ser sinónimos, antónimos, hiperónimos o hipónimos.

2.4. Plagio de opiniones

El plagio no solamente afecta a las empresas sino también a los consumidores. 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 éstos 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. Además, esto puede conllevar un beneficio económico, puesto que cuanto más visitas se consiguen mayores serán los beneficios por publicidad.

3. Análisis de opiniones

En nuestra sociedad interconectada, saturada de mensajes comerciales, conseguir la atención y la credibilidad del potencial resulta cada vez más costoso y difícil. El consumidor recurre a la Web en busca de opiniones sobre productos y marcas, en las que él mismo puede participar activamente. El deseo de compartir experiencias con marcas y productos es quizá la principal característica de estas nuevas redes sociales.

Diversos estudios demuestran la influencia de la Web 2.0 en las prácticas de consumo: como el estudio realizado por la Asociación para la Investigación de Medios de Comunicación (AIMC[4]), en el que se afirma que el 75.5 % de internautas españoles admite haberse documentado en internet durante el

[1] http://plagiarism.phys.virginia.edu/
[2] http://pan.webis.de/
[3] http://homepages.feis.herts.ac.uk/∼pdgroup/
[4] http://www.aimc.es/aimc.php

último año, como paso previo a formalizar una compra de productos o servicios.

Es por ello que las empresas tienen la obligación de supervisar en los medios sociales las opiniones relacionadas con sus productos y servicios. 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 se convierta en un trabajo completamente irrealizable. Por este motivo 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.

3.1. Anáisis de opiniones basado en ontologías

En un documento donde un cliente opina sobre un producto o servicio, se escriben tanto aspectos positivos como negativos del objeto, aunque el sentimiento general del objeto puede ser positivo o negativo.

Las empresas deben analizar tanto la orientación general de la opinión, así como la orientación de cada concepto del que se opina en el documento evaluativo. Por ejemplo, una empresa de turismo que ofrece un viaje a París, con el hotel *Parisino* incluido, y entradas al museo del *Louvre*; aparecerán opiniones como: *El hotel "Parisino" era desastroso; pero el museo de Louvre era precioso.* 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 *museo* 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 el museo. En el caso que la mayoría tengan la misma opinión, la empresa podría dejar de ofrecer el viaje a París. Sin embargo, analizando las orientaciones semánticas de los conceptos, 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.

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 *museo*, podría extraer los adjetivos de cada concepto y a partir de éstos calcular la polaridad promedio de cada uno de los conceptos.

3.2. Integración de opiniones vía fusión de ontologías

Sin embargo, dado el coste de conseguir la opinión de los consumidores, varias empresas podrían decidir compartir e intercambiar la información que poseen sobre las opiniones de los consumidores. En estos casos, se debe 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. 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.

El algoritmo (Mascardi, Locoro, y Rosso, 2009) propone que la empresa e_1 obtenga la polaridad de los conceptos de su ontología O_1 del conjunto de opiniones que tenga en su base de datos, del mismo modo la empresa e_2 obtendrá la polaridad de los conceptos de su ontología O_2 del conjunto de opiniones que posee en su base de datos. Para la obtención de la polaridad de los conceptos y propiedades de las ontologías cada empresa seguirá los siguientes pasos:

- Se buscan las frases de cada opinión que contienen algún concepto de la ontología de la empresa;
- Seguidamente, se extraen de las frases obtenidas en el paso anterior, los adjetivos adyacentes de cada concepto.
- En el siguiente paso se obtienen la polaridad de los adjetivos utilizando SentiWordNet.

Ontología	Corpus Dividido		Corpus completo	
	Num.	Res.	Num.	Res.
ETP Tourism	1.500	72,41 %	3.000	72,2 %
qallme-tourism	1.500	70,92 %	3.000	71,2 %
Ontology matching	3.000	71,13 %	3.000	71,33 %

Tabla 2: Resultados de los experimentos

- Se comprueba que la frase es afirmativa, en caso contrario, se invierte la polaridad que nos devuelve SentiWordNet.

Posteriormente se realizará una fusión de ontologías mediante una ontología general O (*upper ontology*) y a través de ésta, se realizará un cálculo de la orientación semántica de una opinión t como la suma de las polaridades de cada concepto de la ontología general O.

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 2 se muestran los resultados obtenidos. Un dato destacable es que tras realizar el proceso de fusión de ontologías se obtienen resultados muy cercanos a los resultados 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 obtenidos 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 a la fusión de ontologías.

4. Conclusiones

4.1. 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. Una empresa debe proteger su material intelectual, pues su mayor éxito en el mercado son sus productos o servicios que la diferencian del resto de empresas.

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 y sus ideas publicadas en la Web. La idea era investigar hasta qué punto se podría hacer utilizando el software de detección de plagio que se encuentra disponible en la Web. Los pobres resultados que obtuvimos con la herramienta WCopyFind, así como con Ferret, nos han demostrado la necesidad de desarrollar métodos de detección automática de plagio para empresas.

4.2. Cómo beneficiarse de las ventajas de la Web 2.0

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. Una de las informaciones más importantes que se encuentra hoy en día en la Web 2.0 son las opiniones de los consumidores sobre los productos y servicios de las marcas. Esta información ayuda a las empresas a detectar las tendencias del mercado. Por ello, varias empresas pueden decidir compartir los análisis de opiniones. En este trabajo, hemos comprobado como al realizar la integración de las opiniones vía fusión de ontologías no se pierden datos de los anteriormente calculados por el análisis de opiniones.

Bibliografía

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

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

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

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