Agrupamiento de textos cortos en dominios cruzados

Alba Núñez-Reyes[1], Erick Monroy-Cuevas[1], Esaú Villatoro-Tello[2], Gabriela Ramírez-de-la-Rosa[2], Christian Sánchez-Sánchez[2]

[1] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, Maestría en Diseño, Información y Comunicación (MADIC), División de Ciencias de la Comunicación y Diseño, México

[2] Universidad Autónoma Metropolitana (UAM) Unidad Cuajimalpa, Departamento de Tecnologías de la Información, México

{ar.nunezreyes, ermoncu}@gmail.com, {gramirez,evillatoro,csanchez}@correo.cua.uam.mx

Resumen. Recientemente, las redes sociales se han vuelto un medio ideal para compartir información en tiempo real. Diversos tipos de usuarios las emplean para comentar sus gustos, actividades u opiniones. Esta información vertida en estos medios se ha vuelto de particular interés para los analistas de reputación en línea, pues a través de éstas logran identificar tendencias relevantes. Sin embargo, analizar miles de datos se vuelve una tarea tediosa para el humano. Existen técnicas de clasificación de documentos las cuales representan soluciones alternativas al problema anterior, pero dada la dinámica de las redes sociales, el tener un modelo de clasificación de documentos por tendencias se vuelve una tarea impensable, pues al surgir nuevas temáticas o al cambiar de dominio de análisis, los modelos construidos no se desempeñarán eficientemente. En este trabajo presentamos un método no supervisado para la identificación de temáticas en textos cortos. Nuestros resultados experimentales muestran que el método propuesto permite tener una representación de textos robusta, que se comporta satisfactoriamente en dominios cruzados.

Palabras clave: Agrupamiento de documentos, dominios cruzados, representación de documentos, selección de atributos, coeficiente de Silhouette.

Cross-domain Clustering for Short Texts

Abstract. Nowadays, social networks have become an ideal tool for sharing information in real time. Different type of users use social media to comment about their activities, opinions, personal views, etc. The information poured in this media has become of particular interest to online reputation analysts, for instance, to identify relevant tendencies.

However, the analysis of great amount of data is a very tedious task for a human. There are classification techniques that present alternative solutions for this problem, but given the dynamism of these social networks, having a model for each trend is not feasible since every day are emerging new trends, and even worse, new trends in new domains. In this paper, we present an unsupervised method for short texts categorization. Our experimental results show that our proposed method allows a robust text representation that performs well in cross-domains problems.

Keywords: Document clustering, cross-domain, document representation, feature selection, Silhouette coefficient.

1. Introducción

Actualmente algunos esfuerzos de investigación en materia de Procesamiento del Lenguaje Natural se han enfocado en la aplicación de técnicas de agrupamiento de datos para identificar las temáticas de documentos cortos, en particular de tuits [13]. Twitter se ha convertido en una herramienta de micro-blogging que permite dar seguimiento, en tiempo real, a gran diversidad de eventos que suceden alrededor del mundo. La información vertida en esta red social se ha vuelto de interés para distintos sectores, *e.g.*, económico, académico, político, comercial y empresarial; pues la información contenida en los 140 caracteres permiten a éstos posicionar opiniones o definir tendencias.

El Analista de Reputación en Línea (ARL) es la persona encargada de analizar opiniones e identificar tendencias respecto a una figura pública y/o entidad de interés. Una vez que el ARL logra identificar información que le es relevante, es capaz de sugerir estrategias de mercado, de tal manera que pueda invertir o potenciar determinadas posturas. El principal problema que enfrenta la figura del ARL es la imperiosa necesidad de analizar miles de textos (tuits) para poder realizar de manera eficiente su labor. Dado esto, en el año 2012 se propone por primera vez un marco de evaluación, llamado RepLab [3,1,2], cuyo principal objetivo es impulsar el desarrollo de sistemas automáticos que apoyen en las actividades de un ARL.

Entre los retos propuestos en el RepLab, se propuso la tarea de identificación de tópicos (*topic detection*), la cual consiste en desarrollar métodos que sean capaces de agrupar tuits relacionados por un tópico en común, con el objetivo de permitir al ARL obtener conjuntos de tuits que refieren a la misma temática. En general, el agrupamiento de datos se define como la tarea de construir grupos de objetos, de tal manera que los elementos de un mismo grupo sean muy similares entre si, pero diferentes a los elementos de otro grupo [6], propiedades conocidas como homogeneidad y heterogeneidad respectivamente.

En el área de Procesamiento de Lenguaje Natural se han propuesto diversas estrategias de agrupamiento que funcionan muy bien cuando se trata de documentos formales (*e.g.*, artículos, noticias, libros, etc.) y que además

emplean técnicas de aprendizaje supervisadas, es decir, tienen un conjunto de entrenamiento etiquetado (una clase por tópico) que permite la construcción de modelos de clasificación confiables [14]. Sin embargo, en un contexto en el que los tópicos son muy diversos, e incluso desconocidos, el contar con datos etiquetados se vuelve una práctica muy costosa y frecuentemente inimaginable.

Con la finalidad de resolver los problemas mencionados, en este trabajo proponemos un método de clasificación no supervisado (*i.e.*, agrupamiento) en el cual se empleó una forma de representación compacta de los textos. Nuestra hipótesis es que la alta dimensionalidad de técnicas tradicionales de representación de documentos afecta el comportamiento de algoritmos de agrupamiento cuando se trata de documentos cortos y además de dominios distintos. En este sentido, una representación compacta, la cual elimina términos muy especializados (dependientes del dominio) y términos muy comunes, permitirá generar un agrupamiento temático más confiable para dominios cruzados.

Para realizar nuestros experimentos trabajamos con una muestra de los datos proporcionados por los organizadores del RepLab del 2013[3]. Los resultados obtenidos muestran que la forma de representación propuesta para realizar el agrupamiento de dominio cruzado, permite obtener un comportamiento similar al que se logra cuando se trabaja en un esquema de "in-domain", es decir, cuando se evalúa el método en el mismo dominio en el que fue construida la representación.

El resto de este documento se encuentra organizado de la siguiente manera. En la sección 2 se describen algunos de los trabajos relacionados al problema de agrupamiento de textos cortos. En la sección 3 se describe en detalle nuestro método propuesto para el agrupamiento de documentos en un escenario de dominio cruzado. En la sección 4 se describe la metodología experimental y los resultados obtenidos. Finalmente, en la sección 5 se plantean las conclusiones obtenidas y se describen algunas lineas de trabajo futuro.

2. Trabajo relacionado

El reto al momento de trabajar con documentos cortos es, principalmente, que la estructura de los textos cortos no sigue las convenciones léxicas y sintácticas de la mayoría de documentos formales, razón por lo cual se presentan dificultades con métodos tradicionales de agrupamiento. En consecuencia, la forma de representación de textos cortos se ha convertido en un área de interés para la comunidad científica, sobre todo cuando se quieren proponer representaciones que sean robustas a cambios de dominio, donde se sabe que la distribución de los atributos será distinta entre dominios. Tradicionalmente, la forma de representación de textos empleada por técnicas de agrupamiento es la bolsa de palabras (BOW). Esta consiste en representar a un documento como el conjunto total de las palabras que aparecen en él, no obstante, esta técnica tiende a ignorar las relaciones semánticas entre las palabras, por lo cual se pierde gran parte del significado de un documento.

[3] `http://nlp.uned.es/replab2013/`

En el trabajo descrito en [8] los autores proponen eliminar las limitaciones de la representación BOW por medio de enriquecer el texto haciendo uso de ontologías. La idea intuitiva de este enfoque consiste en identificar relaciones entre conceptos de los artículos de Wikipedia y los términos presentes en la bolsa de palabras de cada documento; una vez identificados, los últimos son considerados para enriquecer el texto del documento en revisión. De manera similar, en [4] se propone un método de agrupamiento de notas extraídas de servicios RSS[4] y snippets[5] de Google, el cual también busca enriquecer la BOW empleando Wikipedia, para tener un agrupamiento más efectivo.

En el trabajo descrito en [10], los autores proponen un método de agrupación de textos cortos por medio de la identificación de términos "núcleo". En esencia, el método propuesto es un proceso iterativo que identifica el término núcleo de un conjunto de textos cortos y realiza un primer agrupamiento. Posteriormente, el proceso de identificación de términos núcleo se repite y genera una nueva propuesta de agrupamiento. Este proceso se repite hasta que un criterio de paro es alcanzado, el cual se basa en una medida de calidad de los grupos formados.

Por otro lado, un trabajo que intenta incorporar información secuencial por medio del uso de n-gramas de palabras es el descrito en [7]. Esta propuesta es evaluada en documentos cortos escritos en Mandarín. Los autores argumentan que esta forma de representación permite capturar, además de información léxica, información sobre la estructura de los textos y algunos aspectos semánticos.

Finalmente, en el trabajo descrito en [16] el problema de agrupamiento en dominio cruzado es aproximado por medio de identificar atributos "crudos" que incorporan, en la representación de los textos tanto aquellos atributos que son compartidos por ambos dominios, como atributos que son completamente disjuntos. Para lograr esta incorporación, los autores utilizan un método probabilístico (EM). A pesar de que este trabajo enfrenta el mismo problema que nosotros queremos resolver, su método requiere de una etapa de entrenamiento, lo cual lo vuelve dependiente de los datos etiquetados disponibles así como dependiente del dominio.

En los trabajos descritos previamente, se pueden identificar las siguientes desventajas: *i)* el problema del agrupamiento en dominio cruzado no es contemplado como problema primario, *ii)* los que enfrentan el problema de dominio cruzado dependen de la existencia de datos etiquetados, y *iii)* la forma de representación propuestas son de muy alta dimensionalidad. Así entonces, en este artículo proponemos usar una forma de representación compacta de los documentos, la cual no considera a elementos dependientes del dominio y al mismo tiempo elimina términos muy comunes. Note que estos dos tipos de atributos pueden ser los causantes de ruido al momento de hacer el agrupamiento.

[4] Siglas de Really Simple Syndication, un formato XML para compartir contenido en la web.

[5] Término extraído del idioma Inglés que refiere a pequeños fragmentos de texto.

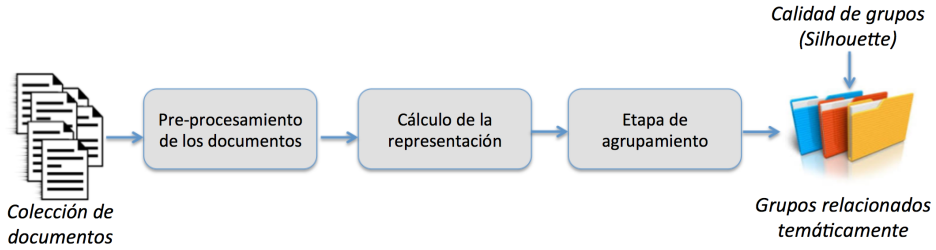

Fig. 1. Arquitectura general de un método de agrupamiento

3. Método propuesto

Como se ha mencionado antes, en el caso del agrupamiento como tarea no supervisada, el problema es encontrar y formar grupos significativos a partir de una colección de elementos no etiquetados. En cierta forma, las etiquetas son asociadas con el número de grupos, la característica de estas etiquetas es que son deducidas únicamente a partir de los elementos de entrada (Figura 1).

Como se puede observar en la Figura 1, los pasos involucrados en la tarea de agrupamiento de documentos son: *1)* pre-procesamiento de la información, esto es, la eliminación de los elementos que podrían causar ruido al momento de realizar el agrupamiento, por ejemplo etiquetas HTML, palabras funcionales, números, signos de puntuación, etc.; *2)* construcción de la representación de los documentos y al mismo tiempo la definición de una medida de proximidad apropiada al conjunto de datos; *3)* aplicación de un algoritmo de agrupamiento; y *4)* la evaluación sobre la calidad de los grupos formados.

Es importante mencionar que en un esquema de agrupamiento de dominio cruzado, el principal objetivo de este trabajo, el reto está en identificar una forma de representación robusta que permita este cambio temático de la colección de documentos entre dominios. Por lo tanto, se busca que se generen grupos de calidad en dominios distintos. Así, nuestro principal interés está en la evaluación de distintas formas de representación para el problema de dominios cruzados. A continuación describimos en detalle cómo se abordaron cada uno de los pasos descritos en la Figura 1.

3.1. Pre-procesamiento de los tuits

Previo al proceso de representación de los documentos se realizó un preprocesamiento a los tuits el cual consistió en los siguientes pasos:

1. Los tuits se convierten a minúsculas con la finalidad de normalizar el vocabulario.
2. Cualquier secuencia de espacios en blanco se convierte en un solo espacio.
3. Se eliminaron las menciones a usuario (*@usuario*) así como cualquier URL que existiera en los tuits.

4. Se eliminan los signos de puntuación. Esto también elimina cualquier emoticono que pudiera aparecer, puesto que no fueron tomados en cuenta para el funcionamiento de este modelo.
5. Cada palabra de un tuit es llevada a su raíz léxica. Este proceso se lleva a cabo mediante el lematizador Porter [11].
6. Se eliminan las palabras vacías y/o funcionales.

3.2. Método base para la representación de los documentos

Una vez eliminados los elementos considerados ruidosos en la etapa de preprocesamiento, el paso obligado es el *indexado* de los documentos de entrenamiento (T), actividad que denota hacer el mapeo de un documento d_j en una forma compacta de su contenido. La representación más comúnmente utilizada para representar cada documento es un vector con términos ponderados como entradas, concepto tomado del modelo de espacio vectorial usado en recuperación de información. Esto es, cada texto d_j es representado como el vector $\overrightarrow{d}_j = \langle w_{kj}, \ldots, w_{|\tau|j} \rangle$, donde τ es el *diccionario*, *i.e.*, el conjunto de términos que ocurren al menos una vez en algún documento de T, mientras que w_{kj} representa la importancia del término t_k dentro del contenido del documento d_j. En ocasiones τ es el resultado de filtrar las palabras del vocabulario, *i.e.,* resultado de un pre-procesamiento (Sección 3.1). Una vez que hemos hecho los filtrados necesarios, el diccionario τ puede definirse de acuerdo a diferentes criterios, sin embargo el que se empleó como método base en esta propuesta corresponde a la Bolsa de Palabras (BOW).

Los pesos w_{kj} pueden ser definidos de variadas formas, sin embargo la que nosotros empleamos es la de ponderado booleano; este pesado consiste en asignar el peso de 1 si la palabra ocurre en el documento y 0 en otro caso. La razón principal para seleccionar un esquema de pesado booleano es debido a la misma naturaleza de los documentos, *i.e.,* tuits. Ya que los tuits son documentos muy cortos (140 caracteres) consideramos que el uso de frecuencias no incorporaría información relevante al algoritmo de agrupamiento.

3.3. Método propuesto para la representación de los documentos

La representación tradicional BOW trae un costo agregado, que es el producir un espacio de términos (atributos) τ de alta dimensionalidad (*i.e.*, $|\tau| \to \infty$). Esta alta dimensionalidad puede ocasionar problemas de *sobre-ajuste* en el proceso de agrupamiento, *i.e.,* el fenómeno por el cual un método de agrupamiento se adapta a las características contingentes de T, en lugar de únicamente a las características constitutivas de las categorías; esto provoca problemas de efectividad pues el algoritmo de agrupamiento tiende a comportarse mejor sobre los datos con los que ha sido evaluado, sin conservar esta tendencia en conjuntos de datos distintos (*i.e.,* dominios diferentes).

Uno de los métodos que ha mostrado ser efectivo como técnica de reducción de dimensionalidad es aquel que conserva sólo los términos que se encuentran

alrededor del punto de transición (pt_T) [5]. El pt_T es un valor de frecuencia que divide a los términos del vocabulario τ en dos conjuntos de términos, los de baja y alta frecuencia [17,5]. En los estudios realizados en [17,5] se demuestra que los términos de frecuencia media están fuertemente relacionados con el contenido de los documentos, lo cual permite resolver efectivamente tareas como el agrupamiento temático de textos. Nuestra hipótesis es que por medio de utilizar esta representación compacta, la cual considera términos de frecuencias medias a altas, es posible generar un modelo de agrupamiento robusto al cambio de dominio. La forma tradicional de calcular el punto de transición es:

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

donde I_1 representa el número de palabras con frecuencia 1 en el vocabulario τ.

3.4. Medida de proximidad

Para todos los experimentos realizados se utilizó a la medida del *coseno* como métrica de proximidad. El objetivo de esta métrica es contar con un valor numérico al cual llamaremos coeficiente de similitud SC, el cual nos dirá cuán parecidos son los documentos d_i y d_j dados. La idea básica de la medida del coseno es determinar el ángulo entre el vector de d_i y de d_j, para hacerlo, calculamos:

$$SC(d_i, d_j) = \frac{\sum_{k=1}^{t} w_{ik} w_{jk}}{\sqrt{\sum_{k=1}^{t} (w_{jk})^2 \sum_{k=1}^{t} (w_{ik})^2}}, \tag{2}$$

donde k va de 1 al número total de términos del vocabulario τ, w_{ik} indica la importancia del término k en el documento d_i mientras que w_{jk} indica la importancia del término k en el documento d_j.

3.5. Algoritmos de agrupamiento

El agrupamiento de documentos consiste en que dado un conjunto de N documento, $\mathcal{D}$, se busca dividirlos o particionarlos en un número predeterminado de k subconjuntos tal que $\mathcal{D} = \bigcup_{i=1}^{k} \mathcal{D}_i$, de tal forma que los documentos asignados a cada subconjunto sean más similares entre sí que a los documentos asignados a otros subconjuntos, *i.e.*, minimizar la distancia intra-clusters o maximizar la semejanza intra-clusters [9].

En este trabajo utilizamos dos métodos de agrupamiento que han mostrado buen desempeño en agrupamiento de documentos de textos: algoritmo de agrupamiento estrella y algoritmo de agrupamiento *k-means*. Ambos algoritmos se describen brevemente a continuación.

Algoritmo estrella. Este es un algoritmo dinámico que induce de manera natural el número de grupos a formar y la estructura de los temas dentro del espacio de textos. Es un algoritmo de tipo particional basado en grafos. Entre sus ventajas respecto a otros algoritmos de agrupamiento están: la no necesidad de conocer la cantidad de grupos a formar, no impone restricciones para la representación de los objetos ni supedita a estos a una medida de semejanza específica. Por otro lado entre sus desventajas podemos mencionar: la dependencia de los grupos obtenidos respecto al orden de análisis de éstos, no permite adherir o eliminar objetos múltiples durante el proceso y obtiene bastantes grupos con pocos elementos [12].

El algoritmo estrella se representa por medio de un grafo de similaridad de la forma $G = (V, E, w)$ donde los vértices (V) representan los textos y cada arista (E) muestra la similaridad entre dos documentos; finalmente el valor del umbral (w) determina el peso de cada arista. La similaridad entre dos documentos se mide con base en alguna medida de similitud como: la cosenoidal, Euclidiana, Jaccard o Manhathan. Las características sobresalientes de este algoritmo son la posibilidad de tener diferentes *estrellas*.

Algoritmo *k-means*. Este algoritmo es uno de los más utilizados para realizar agrupamiento de datos. Esta técnica tiene como objetivo dividir un conjunto de N elementos en un número predeterminado de grupos k [6]. El algoritmo es sencillo y eficiente, pues permite procesar patrones de forma secuencial. La desventaja del k-means radica en que los primeros patrones determinan la configuración inicial de los grupos y su comportamiento depende enormemente del parámetro k. La idea intuitiva es determinar k centros para cada grupo, luego la distancia entre cada centro determinará el resultado del agrupamiento, por tal razón se recomienda ubicarlos lo más alejados posible entre ellos.

3.6. Evaluación del agrupamiento

La medida de validez interna del agrupamiento, que es la que utilizaremos en este trabajo, consiste en determinar dos aspectos del agrupamiento realizado por un algoritmo dado: por un lado qué tan cohesionados están los grupos entre sí, *i.e.*, se busca que los elementos de un mismo grupo se parezcan más entre ellos que con elementos en otros grupos; y por otro lado determina qué tan separados son los elementos de un grupo con respecto a todos los elementos de otros grupos. Un ejemplo de una medida de evaluación de agrupamiento es el coeficiente de Silhouette.

Coeficiente de Silhouette. El coeficiente de Silhouette muestra qué objetos yacen completamente dentro de un grupo y cuáles están en algún sitio entre grupos. Esta medida tiene un rango de [-1, 1], un valor de 1 indica que el documento está lejos de agrupamientos vecinos, 0 indica que el documento está en o muy cerca de la frontera de decisión entre dos grupos vecinos, y valores negativos indican que el documento podría haber sido mal asignado al grupo.

Tabla 1. Estadísticas de los documentos en el corpus empleado

	Dominio origen	Dominio destino
Num. documentos	1000	1000
Promedio de palabras por docs.	9.8	9.9
Promedio de caracteres por docs.	64.25	65.82
Longitud promedio de palabra	9.5	9.62
Diversidad léxica	1	1.02
Vocabulario promedio por docs.	6.1	6.23

Un promedio de Silhuette cercano a 1 indica que los documentos están agrupados correctamente. Mediante los valores de este coeficiente se puede decidir el número de grupos a formar de un conjunto de datos; sin embargo, esto depende en gran medida del método de agrupamiento utilizado. En los experimentos realizados en este trabajo, la medida de *Silhouette* resulta útil para seleccionar el número de grupos apropiado cuando se utiliza el algoritmo *k-means*; por su parte, en el algoritmo estrella, no es posible modificar el número de grupos generados, utilizaremos entonces el coeficiente para validar la efectividad de los agrupamientos y así comparar resultados entre algoritmos.

4. Experimentos y resultados

4.1. Colección de documentos

Para validar nuestra propuesta realizamos experimentos con los datos proporcionados por el RepLab 2013, el cual está formado por aproximadamente 142,000 tuits, tanto en Inglés como en Español. Este corpus está dividido en cuatro grandes dominios: autos, bancos, universidades y música. Cada dominio contiene diferente número de entidades[6]. Para la recolección de este corpus se realizaron búsquedas en Twitter utilizando el nombre del dominio como parámetro de búsqueda, entre el 1 de Junio de 2012 y el 31 de Diciembre de 2012. Cada una de las 61 entidades consideradas en los cuatro dominios tienen alrededor de 2,200 tuits [1]. Para probar nuestra hipótesis se usaron 1000 tuits de los dominios *autos* y *universidades* (500 por dominio), este conjunto de tuits fue considerado para ajustar los parámetros de los algoritmos de agrupamiento. Posteriormente se usaron otros 1000 tuits de los dominios *bancos* y *música* (500 por dominio) para evaluar el agrupamiento en un enfoque de dominio cruzado.

En la Tabla 1 se muestran algunas estadísticas sobre el conjunto de datos utilizados, el conjunto de documentos del dominio origen corresponden a los dominios del RepLab autos y universidades, mientras que el conjunto de documentos del dominio destino corresponden a los dominios del RepLab: bancos y música. En la tabla se pueden observar estadísticas similares para los dos conjuntos de datos, por lo que la diferencia principal radicará en los temas de los documentos de cada dominio.

[6] Una entidad puede ser una figura pública, empresa, institución o un producto.

Tabla 2. Valor del coeficiente de Silhouette (CS) para agrupamiento del algoritmo k-means para 11 valores distintos de k

k:	10	20	30	40	50	60	70	80	100	150	180
BOW	0.05	0.03	0.02	-0.02	0.04	0.09	0.10	0.05	0.02	0.29	**0.31**
pt_T	0.04	-0.13	-0.03	0.00	0.09	0.06	0.05	0.07	0.17	**0.32**	0.24

Tabla 3. Valor del coeficiente de Silhouette (CS) para agrupamiento del algoritmo estrella aplicado a los dominios destino, bancos y música

	Núm. de grupos	Coeficiente de Silhouette
BOW	36	0.03
pt_T	499	0.16

4.2. Resultados de agrupamiento para el dominio origen

Para comprobar la hipótesis planteada en este trabajo, se realizaron experimentos con dos tipos de representaciones, una representación de alta dimensionalidad obtenida mediante la Bolsa de Palabras (BOW) y una representación de baja dimensionalidad obtenida mediante el método de Punto de Transición (pt_T).

Por lo tanto, el primer experimento consiste en utilizar el algoritmo k-means (con distintos valores de k) con ambas representaciones. Los resultados de este experimento se muestran en la Tabla 2. En la tabla se observa que el mejor agrupamiento para ambas representaciones se obtiene para $k = 150$ y $k = 180$, respectivamente. Otro aspecto a notar en la tabla es que cuando se utilizan la representación de BOW, con 2429 términos, se requieren 180 grupos mientras que cuando se utiliza la representación basada en el Punto de Transición (pt_T), con 674 términos, se requieren 150 grupos para obtener resultados ligeramente mejores. Note para todos los casos se buscan valores del coeficiente de Silhouette cercanos a 1 (vea la subsección 3.6).

Durante los experimentos la variación del valor de k llegó hasta 499, para la cual el valor de Silhouette fue prácticamente 1; sin embargo, para los intereses de esta investigación, esto no fue considerado como un resultado deseable para realizar la validación de dominio cruzado, pues prácticamente está considerando dos documentos por grupo.

Para el segundo experimento, se evaluó el desempeño del algoritmo estrella con las mismas dos representaciones. Los resultados de este experimento se muestran en la Tabla 3. En general el desempeño del algoritmo estrella no es adecuado para este conjunto de datos pues aunque pareciera que el agrupamiento con la representación de Punto de Transición es mejor, el número de grupos que se forman son casi la mitad del número total de documentos, por lo que el agrupamiento es demasiado especializado para este corpus.

Tabla 4. Valor del coeficiente de Silhouette (CS) para agrupamiento del algoritmo k-means para 11 valores distintos de k

k:	10	20	30	40	50	60	70	80	100	150	180
pt_T	0.05	0.07	0.08	0.04	0.02	0.05	0.13	0.02	0.17	0.30	**0.34**

De los dos experimentos anteriores se puede concluir que la mejor configuración resulta de utilizar el algoritmo k-means con $k = 150$ y la representación de baja dimensionalidad, esto es, utilizando el método de reducción de dimensionalidad del Punto de Transición.

4.3. Resultados de agrupamiento para el dominio destino

La mejor configuración de parámetros obtenida en el conjunto de documentos del dominio origen se utilizaron para agrupar el conjunto de documento del dominio destino. Cabe mencionar que el dominio destino está compuesto por tuits de los dominios del RepLab banco y música.

Para realizar este experimento en dominios cruzados se representó cada documento del dominio destino con el vocabulario obtenido con el Punto de Transición en el dominio origen. Es decir, se representaron los documento con 674 términos. Los resultados de esta evaluación se muestran en la Tabla 4.

Como puede verse en los resultados de dominio cruzado, la evaluación del agrupamiento con 180 grupos usando una representación de baja dimensionalidad es incluso mejor que para el dominio origen, *i.e.*, $CS = 0.34$ vs $CS = 0.32$, para el escenario de dominio cruzado y el escenario in-domain, respectivamente.

5. Conclusiones

En este trabajo hemos presentando una metodología para resolver el problema de identificación de tópicos en documentos cortos, específicamente tuits. El problema que da origen a este trabajo proviene de las actividades cotidianas que debe realizar un Analista de Reputación en Línea, entre las cuales se encuentra la identificación de temáticas relevantes a la entidad y/o figura pública de su interés.

Debido a la dinámica con que se genera información en las redes sociales, el pensar en diseñar esquemas de clasificación supervisados se vuelve una tarea inimaginable, pues representaría un proceso muy costoso, mismo que eventualmente se volvería obsoleto al poco tiempo debido a que en este tipo de ambientes (Twitter) las temáticas son muy diversas y constantemente cambiantes.

Así entonces, el trabajo realizado se orientó a tratar de eliminar las limitaciones de los esquemas tradicionales, para lo cual se emplearon estrategias de clasificación no supervisadas. Por otro lado, con la finalidad de construir representaciones robustas a distintos dominios, se utilizó una forma de representación compacta de los documentos. La representación empleada permite eliminar términos muy especializados (dependientes del dominio) así como términos

muy comunes (ruido). Los resultados obtenidos son alentadores pues sugieren la posibilidad de llevar a cabo el agrupamiento de este tipo de textos, en dominios cruzados, con resultados satisfactorios. Durante la realización de los experimentos, fue posible notar que los valores obtenidos del coeficiente de Silhouette para el agrupamiento k-means bajo una representación de punto de transición, comprueba la factibilidad de utilizar este tipo de representaciones en dominios diferentes al que fue empleado para la construcción de la representación inicial.

Como trabajo futuro nos interesa evaluar el desempeño del algoritmo de agrupamiento estrella empleando estrategias diferentes para la definición del umbral de similitud, así como métricas distintas, por ejemplo la la similitud suave de coseno, descrita en [15]. En lo que respecta a los algoritmos de agrupamiento, contrario al método de k-means, el algoritmo estrella no requiere de especificar un número de grupos que se quieren formar, por lo cual creemos que si se logra definir un esquema apropiado para determinar el valor del umbral, el comportamiento de éste será más satisfactorio. Agregado a esto, nos interesa igualmente evaluar la propuesta con una muestra mayor de datos, e incluso incorporando información de los otros años del RepLab.

Agradecimientos. Este trabajo fue parcialmente financiado por el CONACyT a través de las becas 587804 y 588090, y el programa del SNI. También se agradece el apoyo otorgado a través de la Coordinación de la Maestría en Diseño, Información y Comunicación (MADIC) de la UAM Cuajimalpa.

Referencias

1. Amigó, E., Carrillo de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Martín, T., Meij, E., Rijke, M., Spina, D.: Information Access Evaluation. Multilinguality, Multimodality, and Visualization: 4th International Conference of the CLEF Initiative, CLEF 2013, Valencia, Spain, September 23-26, 2013. Proceedings, pp. 333–352. Springer Berlin Heidelberg, Berlin, Heidelberg (2013), http://dx.doi.org/10.1007/978-3-642-40802-1_31
2. Amigó, E., Carrillo-de Albornoz, J., Chugur, I., Corujo, A., Gonzalo, J., Meij, E., Rijke, M., Spina, D.: Information Access Evaluation. Multilinguality, Multimodality, and Interaction: 5th International Conference of the CLEF Initiative, CLEF 2014, Sheffield, UK, September 15-18, 2014. Proceedings, chap. Overview of RepLab 2014: Author Profiling and Reputation Dimensions for Online Reputation Management, pp. 307–322. Springer International Publishing, Cham (2014), http://dx.doi.org/10.1007/978-3-319-11382-1_24
3. Amigó, E., Corujo, A., Gonzalo, J., Meij, E., de Rijke, M.: Overview of RepLab 2012: Evaluating online reputation management systems. (2012)
4. Banerjee, S., Ramanathan, K., Gupta, A.: Clustering short texts using wikipedia. In: Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. pp. 787–788. SIGIR '07, ACM, New York, NY, USA (2007), http://doi.acm.org/10.1145/1277741.1277909
5. Booth, A.D.: A "law of occurrences for words of low frequency. Information and Control 10(4), 386 – 393 (1967), http://www.sciencedirect.com/science/article/pii/S001999586790201X

6. Gan, G., Ma, C., Wu, J.: Data clustering: theory, algorithms, and applications, vol. 20. Siam (2007)
7. He, H., Chen, B., Xu, W., Guo, J.: Short text feature extraction and clustering for web topic mining. In: Semantics, Knowledge and Grid, Third International Conference on. pp. 382–385 (Oct 2007)
8. Hu, X., Zhang, X., Lu, C., Park, E.K., Zhou, X.: Exploiting wikipedia as external knowledge for document clustering. In: Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. pp. 389–396. KDD '09, ACM, New York, NY, USA (2009), http://doi.acm.org/10.1145/1557019.1557066
9. Mateos Sánchez, M., García-Figuerola Paniagua, C.: Aplicación de técnicas de clustering en la recuperación de información web. (Biblioteconomía y administración cultural ; 205), Gijón (Asturias) : Trea (2009), http://www.mcu.es/ccbae/es/consulta/registro.cmd?id=173815
10. Ni, X., Quan, X., Lu, Z., Wenyin, L., Hua, B.: Short text clustering by finding core terms. Knowl. Inf. Syst. 27(3), 345–365 (Jun 2011), http://dx.doi.org/10.1007/s10115-010-0299-7
11. Porter, M.F.: An algorithm for suffix stripping. Program 14(3), 130–137 (1980)
12. Pérez Suárez, A., Martínez Trinidad, J., Medina Pagola, J., Carrasco Ochoa, A.: Algoritmos dinámicos para el agrupamiento con traslape. Tech. Rep. CCC-10-001, Instituto Nacional de Astrofísica, Óptica y Electrónica (2010)
13. Rosa, K.D., Shah, R., Lin, B., Gershman, A., Frederking, R.: Topical clustering of tweets. Proceedings of the ACM SIGIR: SWSM (2011)
14. Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. 34(1), 1–47 (Mar 2002), http://doi.acm.org/10.1145/505282.505283
15. Sidorov, G., Gelbukh, A., Gómez-Adorno, H., Pinto, D.: Soft similarity and soft cosine measure: Similarity of features in vector space model. Computación y Sistemas 18(3), 491–504 (2014)
16. Zhuang, F., Luo, P., Yin, P., He, Q., Shi, Z.: Concept learning for cross-domain text classification: A general probabilistic framework. In: Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence. pp. 1960–1966. IJCAI '13, AAAI Press (2013), http://dl.acm.org/citation.cfm?id=2540128.2540409
17. Zipf, G.: Human behaviour and the principle of least-effort. Addison-Wesley, Cambridge, MA (1949), /brokenurl#http://publication.wilsonwong.me/load.php?id=233281783