Tópicos en Base de Datos
Detección de patrones espacio-temporal en Twitter

Universidad Nacional De San Agustín

Juan José López Condori, *Estudiante, UNSA, juanjolopez28@gmail.com*

***Resumen*—Las redes sociales se presentan como valiosas fuentes de información sobre sus usuarios y sus respectivos comportamientos e intereses. Muchos investigadores en minería de datos han analizado estos tipos de datos, con el objetivo de encontrar patrones. En este trabajo se aborda el problema de identificar y mostrar perfiles de tweet mediante el análisis de múltiples tipos de datos: espacial, temporal, social y de contenido,acompañada de una visualización multidimensional . El proceso de minería de datos que extrae los patrones está compuesto por la manipulación de las matrices de disimilitud para cada tipo de datos, que se alimentan a un algoritmo de agrupamiento para obtener los patrones deseados. Este trabajo estudia las funciones de distancia apropiadas para los diferentes tipos de datos, Los métodos de normalización,de reducción de dimensionalidad y los algoritmos de agrupación existentes. La visualización está diseñada para un uso dinámico e intuitivo,con el objetivo de revelar los perfiles extraídos de forma comprensible.**

I. Introducción

En los últimos años, los servicios de redes sociales han alcanzado una gran importancia en la vida social y también en las estrategias empresariales para las empresas, ya que son una fuente oportuna y rentable de recursos espacio-temporales y conductual. La adhesión masiva y el número de plataformas que proporcionan interacción social conduce a un crecimiento en los datos almacenados dentro de estos servicios.

Twitter ha demostrado ser una fuente de datos popular dentro de las redes sociales. Debido al gran número de usuarios activos y al fácil acceso a su API pública [10].Los datos de Twitter probablemente se pueden organizar en subgrupos que representan los perfiles de los tweets y, por tanto, de los usuarios. Estos perfiles pueden ser útiles para muchas tareas (mercadeo, Ciencia política, gobierno, desarrollo de productos, etc.). Sin embargo, dada la cantidad de datos, así como su compleja naturaleza (espacio, tiempo, contenido y social), estos patrones no se extraen fácilmente utilizando estrategias clásicas de minería de datos. Una solución es la capacidad de asignar una importancia diferente a cada dimensión de datos puede resultar útil para controlar el proceso de minería de datos y revelar además patrones ocultos.

Por lo tanto, nuestro objetivo es mostrar una visualización intuitiva con el fin de revelar los perfiles extraídos de forma comprensible y mostrar los patrones minados,ademas de un proceso de minería de datos que permite una combinación ponderada de agrupaciones múltiples en varias dimensiones(espacial, temporal ,contenido y social).

Este artículo está organizado de la siguiente forma: La Sección II se hace una explicación de las medidas de distancia utilizadas que se usaran al momento analizar cada dimensión . En la sección III, se presenta el proceso de combinación de las 4 dimensiones descritas, es decir *Clustering en dimensiones múltiples*. La Sección IV presenta las herramientas de visualización que se usaron para cada una de las dimensiones y finalmente, en la Sección V se discuten los resultados obtenidos.

II. Metricas

Para cada una de las demensiones:Espacial,Temporal,Contenido y la social se aplicaron distintas metricas de distancias,Consideramos que cada tweet se define formalmente como t_i donde i es el identificador del índice en la recopilación de datos de tweet. Las funciones de distancia entre dos tweets t_i y t_j se definen como $Dist^X(ti, tj)$, donde X es la dimensión en la que la función mapea los valores. X puede tomar los valores Sp, T, C, So que son relacionados respectivamente a las dimensiones espaciales, temporales, de contenido y sociales.[1]

II-A. Dimensión Espacial

En esta dimensión se usó la Distancia de Harvesine que es la aproximación esférica entre dos puntos de la superficie terrestre.
Para cada par de tweets t_i y t_j, la función de distancia usa las latitudes ϕ_{ti} Y ϕ_{tj} y longitudes λ_{ti} Y λ_{tj} Para determinar la distancia. El valor R es el radio de la tierra:

$$dist^{Sp}(t_i,t_j) = 2Rsin^{-1}\left(\left[sin^2\left(\frac{\phi_{t_i}-\phi_{t_j}}{2}\right) + cos\phi_{t_i}cos\phi_{t_j}sin^2\left(\frac{\lambda_{t_i}-\lambda_{t_j}}{2}\right)\right]^{0.5}\right)$$

Figura 1: Distancia de Harvesine

II-B. Dimensión Temporal

En cuanto a la dimensión temporal, contrariamente a la dimensión espacial, que se asigna en R_2 , El tiempo está representado En R, la distancia temporal se calculó simplemente como la diferencia De las marcas de tiempo
Para cada par de tweets t_i y t_j, la marca de tiempo Los valores Δi y Δj se utilizan para calcular la distancia:

$$dist^T(t_i, t_j) = |\Delta_i - \Delta_j|$$

Figura 2: Diferencia en Segundos

II-C. Dimensión de Contenido

Para calcular la similitud entre dos textos, hay que explorar las funciones de distancia de minería de texto como la distancia de similitud de coseno.Para aplicar una función de distancia en el texto, las representaciones de documentos deben especificarse en una etapa anterior. La idea principal es crear una matriz de documento-término y extraer los vectores para calcular su disimilitud. El documento principal Las técnicas de representación son TF y TF-IDF. Para un tweet t_i definimos el texto tweet como αi y calculamos su representación TFIDF en una matriz de documentos D

$$TFIDF(\alpha_i, D) = TF(\alpha_i, D) * IDF(\alpha_i)$$

Figura 3: TFIDF

II-D. Dimension Social

En cuanto a la dimensión social,consideramos el peso por la cantidad de retweet de un tweet.En las redes sociales consideramos la importancia de un tweet por la veces que este fue retweeteado basándonos en un estudios realizados en [8] y [9] que se centran en la importancia de los retweets.

III. Clustering en Dimensiones Múltiples

Presentamos en esta sección nuestra conceptos necesarios y nuestra metodología para abordar este problema, incluyendo las funciones de distancia utilizadas, el el algoritmo elegido y la estrategia propuesta para combinar dimensiones.

III-A. Conceptos:

- *Clustering:* Se define como el proceso de agrupar un conjunto de objetos de datos en múltiples grupos o Clusters para que los objetos dentro de un clúster tengan una alta similitud, pero son muy disímiles a los objetos en otros grupos,la evaluación de la semejanza se calcula mediante funciones de distancia [3]
- *Clustering Basado en Densidad:* Los algoritmos basados en densidad localizan zonas de alta densidad separadas por regiones de baja densidad.(Ruido)[4]
- *DBSCAN:* Es un algoritmo de agrupamiento(*Clustering*) basado en densidad porque encuentra un número de grupos (clusters) comenzando por una estimación de la distribución de densidad de los nodos correspondientes. DBSCAN es uno de los algoritmos de agrupamiento más usados y citados en la literatura científica.[5]
- *K-means:* Es un método de agrupamiento, que tiene como objetivo la partición de un conjunto de n observaciones en k grupos en el que cada observación pertenece al grupo cuyo valor medio es más cercano. Es un método utilizado en minería de datos.

III-B. Metodología

La combinación de distancias se refiere al sistema de peso Aplicadas sobre las cuatro dimensiones para Obtener una nueva matriz de disimilitud multidimensional. Los pesos determinan la importancia de las dimensiones, dada como valores porcentuales.
La matriz de disimilitud multidimensional contendrá la suma de todas las disimilitudes Multiplicado por el valor de peso para la dimensión correspondiente, Siempre asegurando que esta suma sea igual al 100 %. Más Formalmente, para cada valor de peso w_{Sp}, w_T, w_C, w_{So}, tenemos que w_{Sp}, w_T, w_C, $w_{So} \in \{0, 0{,}25, 0{,}5, 0{,}75, 1\} \mid w_{Sp}+w_T+w_C+w_{So} = 1$
Sin embargo, para obtener resultados relevantes,es necesario normalizar las matrices de distancia de una sola dimensión, ya que la las escalas para cada dimensión son diferentes. El propósito es usar un una escala similar para todas las dimensiones para asegurar que la importancia de cada dimensión está determinada por los pesos.

IV. Herramientas

Ahora detallaremos las herramientas y las librerias que utilizamos en nuestros Experimentos, así como las recomendadas para la visualización:

Herramienta	Plataforma o Lenguaje	Funcionalidad
API Twitter	Python 2.7	Se utilizo para la recolección de datos de la red social Twitter.Se siguió la documentacion [10]: https://media.readthedocs.org/pdf/python-twitter/latest/python-twitter.pdf
Scikit-learn	Python 2.7	Es una librería sencilla y eficiente para la minería de datos y análisis de datos,se uso para el calculo del TFIDF [11] http://scikitlearn.org/stable/index.html ademas para la reducción de la Dimensionalidad.
Projection Explorer (PEx)	Java	Se usó para clusterizar nuestros vectores caractiríscas de cada dimensión.[12] http://vis.icmc.usp.br/vicg/tool/1/projection-explorer-pex
Plotly	Pagina Web	Se usó para visualizar tanto la dimension espacial como la social. https://plot.ly/
Qgis	Aplicación	Herramienta recomendada para la visualización http://www.qgis.org/es/site/
Matplotlib	Librería Python	Herramienta con la que se logro la visualización multidimensional http://matplotlib.org/1.4.3/mpl_toolkits/mplot3d/index.html

V. Experimentos

Para los experimentos de nuestro proyecto se realizó en una máquina portátil con las siguiente características:

- Procesador: Intel Core™ i3-380M 2.50 GHz
- Memoria RAM: 4GB DDR3
- Disco Duro: 500GB

Nuestra base de datos fue recopilada utilzando la API de twitter [10] , se recopiló 5000 tweets,este conjunto de datos contiene datos de los meses marzo,abril,mayo y junio del 2016,con tweets escritos en su totalidad en inglés y publicado en varios países a nivel mundial.

Los tweets fueron almacenados en una Base de datos en *Postgres*,se establecieron 4 matrices de distancias previamente

precomputadas *matrizespacial*, *matrizcontenido*, *matriztemporal* y *matrizsocial*,utilizando indices para el fácil acceso a nuestra data.

Los vectores característica por cada dimensión tuvieron el siguiente formatos:

- Dimensión espacial: $[longitud, latitud]$
- Dimensión Temporal: $[Segundos(Dia, Mes, \text{Año}, hora)]$
- Dimensión Contenido: $[1, .., 12]$ Un vector de 12 características luego de hacer una reducción de dimensionalidad con [13].
- Dimension Social:$[Retweets]$

V-A. Análisis del vector característica Dimensión Contenido

El procedimiento para para la extracción de el vector característica del contenido de un tweet fue el siguiente:

- Primero se aplico TFIDF para el calculo de vector característica pero este era demasiado extenso.
- Entonces nuestro segundo paso era reducir la dimensionalidad de este aplicando svd(singular value decomposition) [13] a un vector de tamaño 12.

La prueba se hizo sobre la data de 5000 tweets, aplicamos K-means para clusterizar en 5 grupos nuestros datos como se muestra en Figura 1 y en la Figura 2 los mismos clusters pero con el algoritmo Neighbor-joining.

- **Puntos Rojos:** Estos tweets hablan sobre carros y particularmente sobre Tesla Motors Inc.
- **Puntos Amarillos:** Estos tweets hablan sobre farmacéuticos y sobre las compañías que se dedican a estos productos
- **Puntos verdes:** Estos tweets hablan sobre la compra de producto onnline sobre todo en Amazon.
- **Puntos Azules:** Estos tweets hablan sobre dinero centrándose en el digital, en Paypal.
- **Puntos Celestes:** Estos tweets hablan sobre Microsoft.

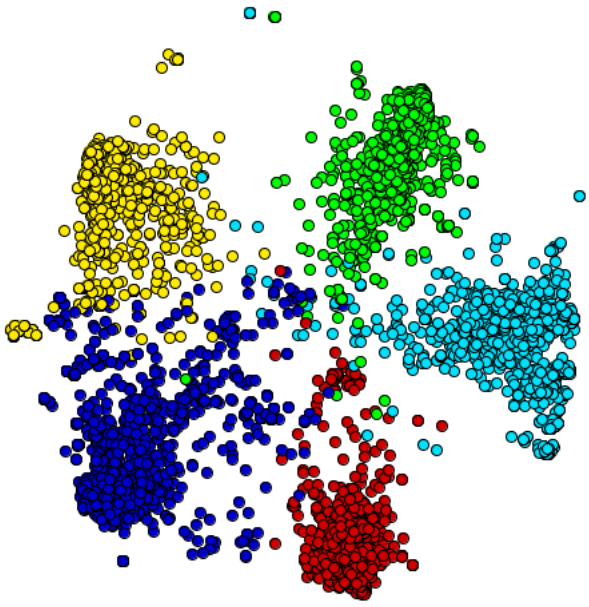

Figura 4: Ploteo y clustering usando k-means(Dimensión de contenido)

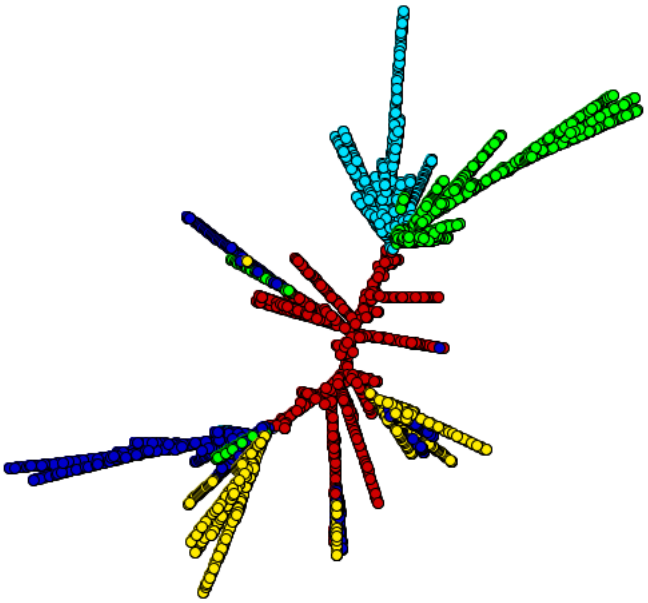

Figura 5: Clustering con neighbor-joining(Dimensión de contenido)

V-B. Análisis del vectores característica multidimensional

Para el análisis de la matriz multidimensional y aplicar nuestro método trabajamos nuestra visualización, para que esta sea intuitiva utlizando la librería *matplotlib* de Python.

En la Figura 6 se muestra el clustering dando 100 % de importancia a la dimensión de contenido se forman 5 clusters,cada uno de estos hablan de temas distintos el color rojo sobre la empresa tesla y sobre automóviles,el azul sobre enfermedades y medicina, el verde sobre amazon,el purpura sobre payla y pagos onnline y el negro sobre microsoft. en la Figura 7 se muestra el clustering con un 100 % de importancia en la dimensión espacial dándonos 3 cluster: America, Europa y Asia junto con Australia. En la Figura 8 dando importancia del 50 % tanto a la dimensión de contenido y espacial vemos 4 clusters que esta bien definidos en cuanto a su contenido y espacialmente vemos que en Australia la mayoría de tweets habla sobre Amazon.
Finalmente en la Figura 9 el ejemplo dando la importancia del 50 % a la dimensión de contenido y temporal respectivamente vemos que en el mes abril y mayo en America del sur la mayoría de tweets hablan sobre Amazon y sobre autos. además de eso por cada cluster se muestra el tweet más importante que se representa por prismas que son proporcionales a la cantidad de retweets que tiene dicho punto.

VI. Conclusiones

Los objetivos de este trabajo fueron utilizar el clustering para identificar patrones a través de dimensiones de datos de naturaleza muy diferente (por ejemplo, espacio y contenido), lo que permite al usuario controlar la importancia relativa de cada uno de ellos en el proceso. Para lograr estas metas, un proceso de minería de datos fue desarrollada con diferentes etapas: preparación de datos, cálculo de las matrices, normalización y combinación. Por último, agrupación utilizando

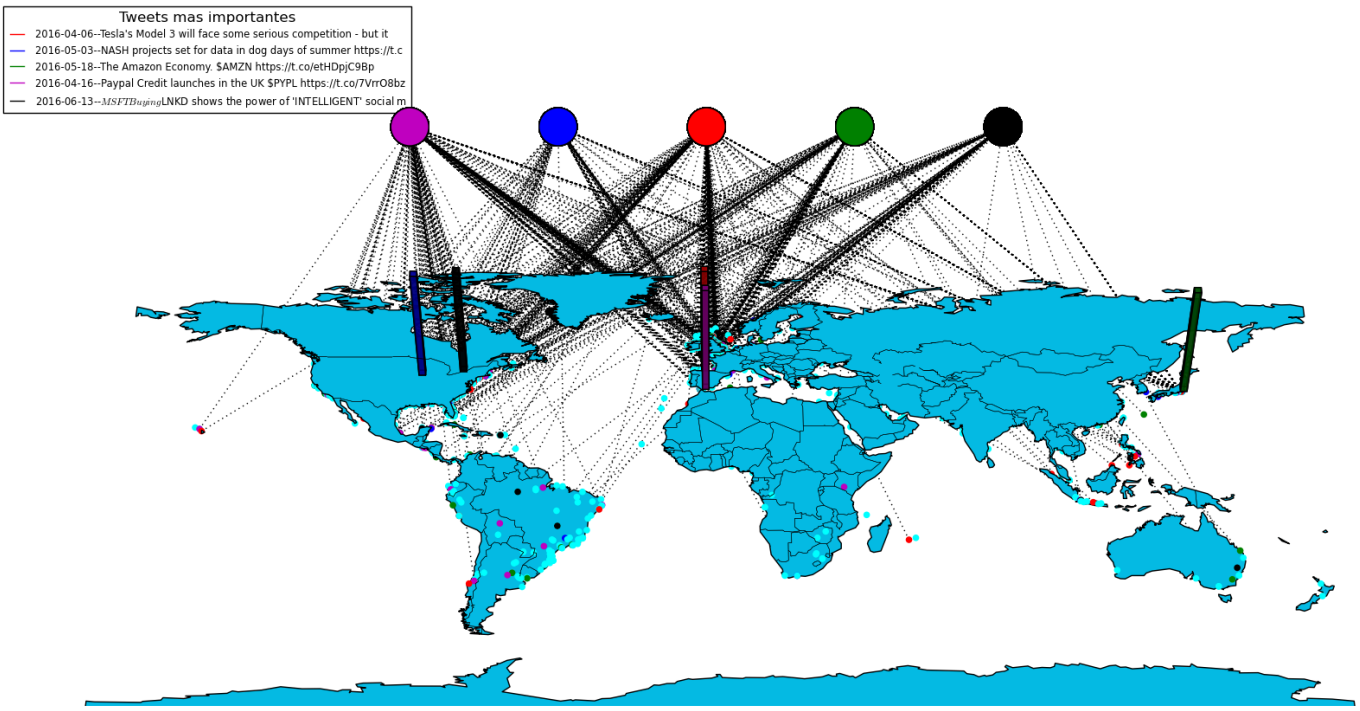

Figura 6: Ejemplo Clustering con 100 % de importancia(peso) asignado a la dimensión de contenido.

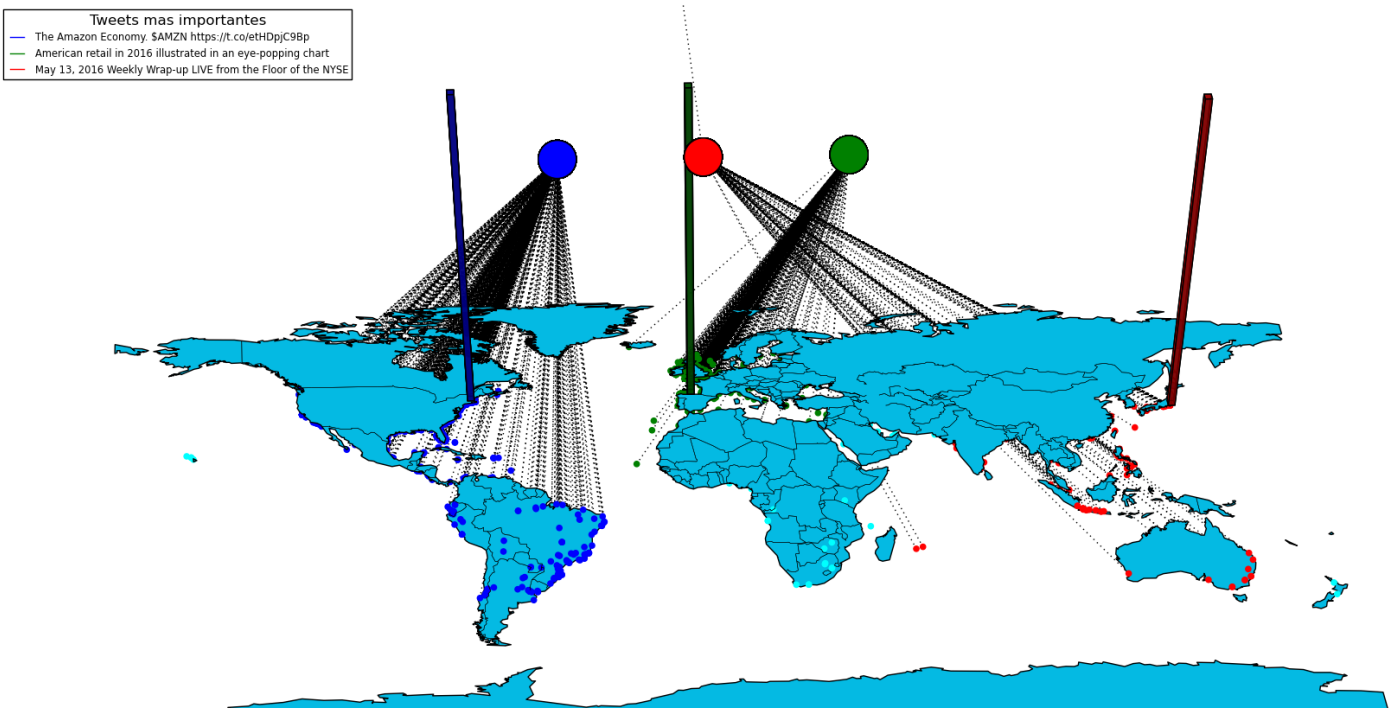

Figura 7: Ejemplo Clustering con 100 % de importancia(peso) asignado a la dimensión espacial.

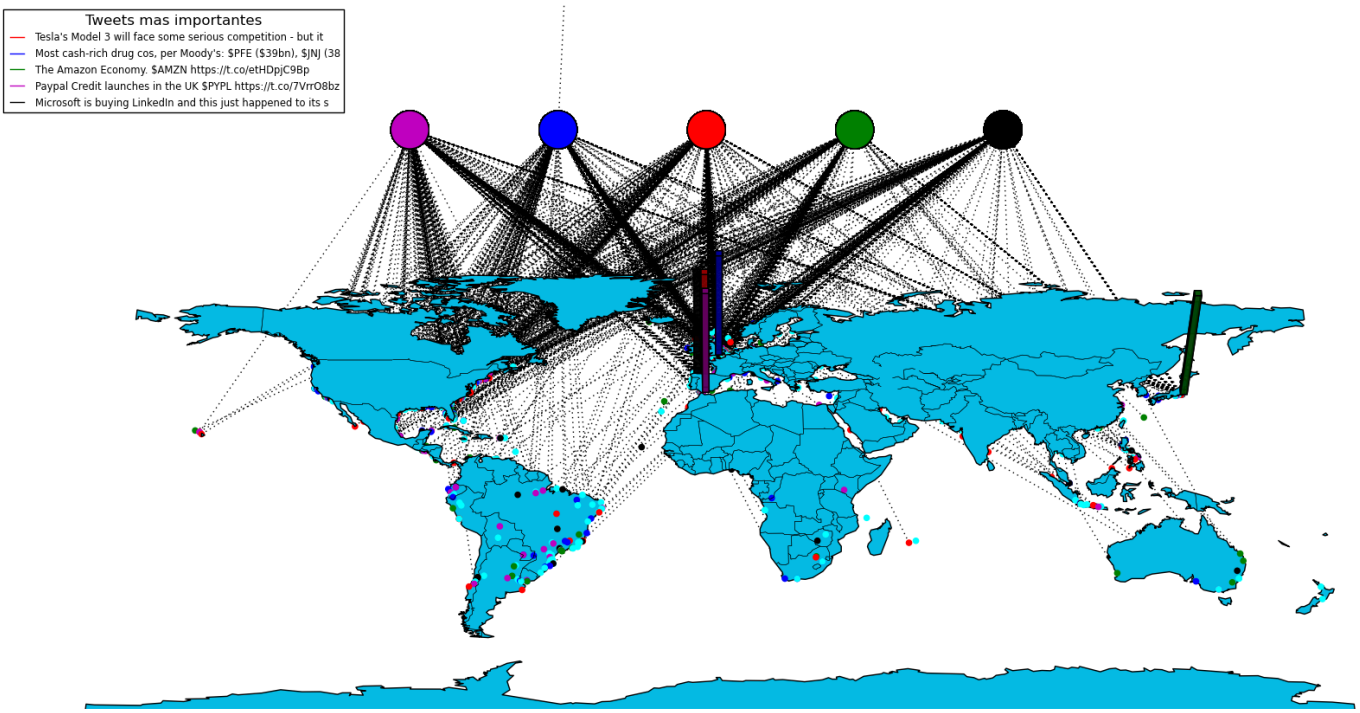

Figura 8: Ejemplo Clustering con 50 % de importancia(peso) asignado a la dimensión de contenido y 50 % a la dimensión espacial

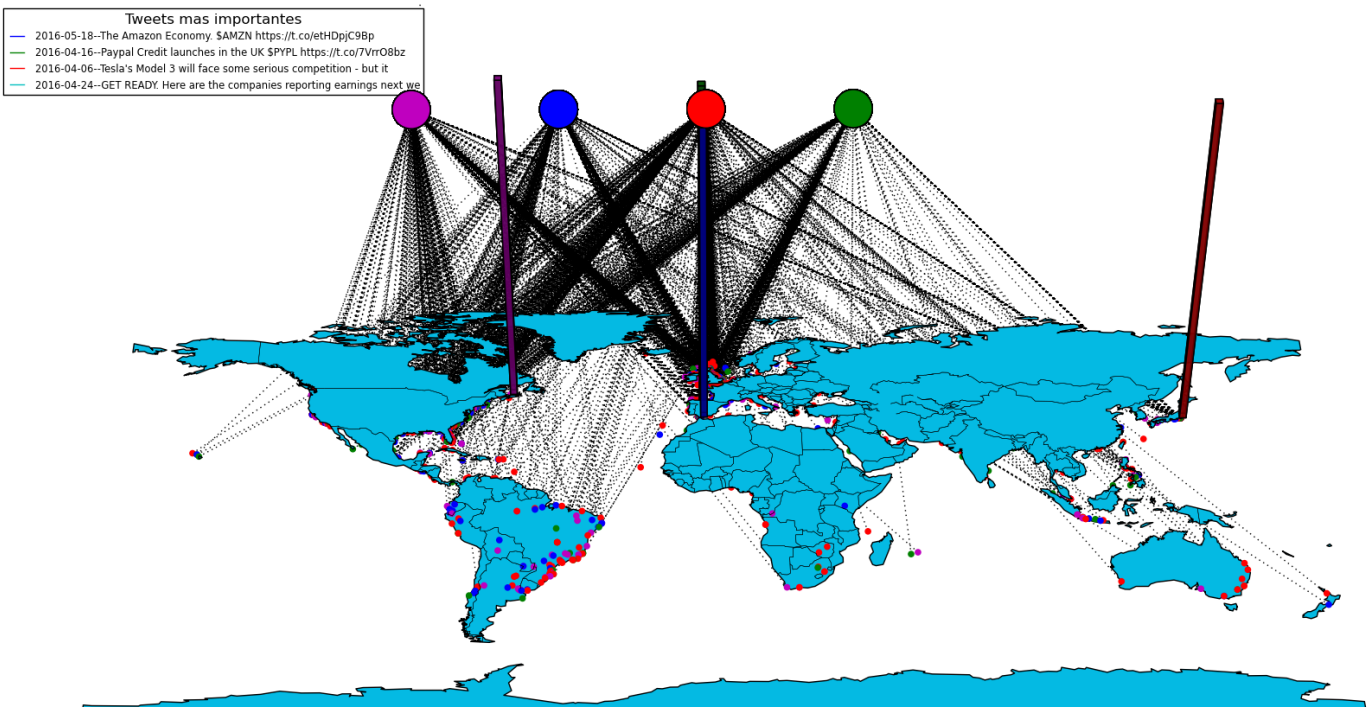

Figura 9: Ejemplo Clustering con 50 % de importancia(peso) asignado a la dimensión temporal y 50 % a la dimensión de contenido.

esas matrices. Se combinaron utilizando un conjunto de pesos predefinidos, representando diferentes niveles de importancia de cada dimensión. Las agrupaciones se obtuvieron ejecutando DBSCAN en estas matrices y ademas de K-means. La herramienta de visualización representa esos patrones. Es decir, un Mapa, un cuadros con los tweets mas importantes por cada cluster y un prisma horizontal respecto al dichos tweets. La primera tarea para el trabajo futuro es estudiar mejores maneras representan las dimensiones sociales y de contenido, para facilitar la interpretación de los resultados y dado que el método propuesto para la agrupación en varios dimensiones es genérico, también se debe evaluar la agrupación con otras dimensiones, como por ejemplo imágenes.

REFERENCIAS

[1] TIAGO CUNHA, CARLOS SOARES,EDUARDA MENDES RODRIGUES,*TweeProfiles: Detection of spatio-temporal patterns on Twitter*,2013.

[2] TIAGO CUNHA *TweeProfiles: detection of spatio-temporal patterns on Twitter*,Facultad de Ingenería de la Universidad de Porto

[3] J. HAN, *Data Mining : Concepts and Techniques (2nd Edition)*,Pag 383-386, 2006

[4] DAMARIS PASCUAL GONZÁLEZ, *Algoritmos de Agrupamiento basados en densidad y Validación de clusters*, Tesis Doctoral,2010

[5] MARTIN ESTER, HANS-PETER KRIEGEL, JIIRG SANDER, XIAOWEI XU,*A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise*,1996

[6] MIHAEL ANKERST, MARKUS M. BREUNIG, HANS-PETER KRIEGEL, JÖRG SANDER,*OPTICS: Ordering Points To Identify the Clustering Structure*,1999

[7] J. HENCIL PETER† ,A.ANTONYSAMY ,*Heterogeneous Density Based Spatial Clustering of Application with Noise* ,2010

[8] RIO AKASAKA,PATRICK GRAFE ,MAKOTO KONDO,*An Analysis of Viral Retweets on the Twittersphere* ,2010

[9] MEENAKSHI NAGARAJAN, HEMANT PUROHIT, AMIT SHETH *A Qualitative Examination of Topical Tweet and Retweet Practices*,2010

[10] PYTHON-TWITTER DOCUMENTATION,https://media.readthedocs.org/pdf/python-twitter/latest/python-twitter.pdf

[11] SCIKIT-LEARN,http://scikit-learn.org/stable/index.html

[12] PROJECTION EXPLORER (PEX),http://vis.icmc.usp.br/vicg/tool/1/projection-explorer-pex

[13] SINGULAR VALUE DECOMPOSITION http://matpalm.com/lsa_via_svd/intro.html

[14] PLOLY https://plot.ly/