Procesamiento del Lenguaje Natural, Revista nº 54, marzo de 2015, pp 53-60 recibido 24-11-14 revisado 27-01-15 aceptado 13-02-15

Anotación y representación temporal de *tweets* multilingües

Temporal annotation and representation of multilingual tweets

Asunción Vázquez-Méndez, Ana García-Serrano
ETSI Informática UNED
C/Juan del Rosal, 16
28040 Madrid
avazquez254@alumno.uned.es, agarcia@lsi.uned.es

Resumen: El tiempo es un elemento de importancia capital en todo espacio de información y Twitter no es una excepción. La explotación de la información temporal en tareas de recuperación y organización de información, tiene una larga tradición. Sin embargo, esta clase de enfoques, basados en contenido, no han sido muy explorados para el dominio de Twitter, y en consecuencia escasean los Corpus de *tweets* anotados con información temporal. En este artículo, se propone un modelo de anotación de la información temporal en el dominio de Twitter, basado en el Análisis de Conceptos Formales, en el que los atributos del contexto serán las expresiones temporales, eventos y tipos de eventos presentes en los *tweets*. Se define un Calendario especialmente adecuado a los fenómenos de conmemoración de aniversarios y fechas señaladas en Twitter, el Calendario Imaginario-Colectivo. El Corpus de estudio ha sido extraido de la colección de RepLab2013. Se incluye un completo análisis del mismo desde una perspectiva temporal.
Palabras clave: Información temporal, Anotación temporal de tweets, Representación de información basada en contenido

Abstract: Time is a crucial element in any space of information and Twitter is no an exception. Although the exploitation of temporal information in retrieval and organization tasks has a long tradition, content-based approaches have not been fully explored for Twitter and researchers lack of sufficient Corpus annotated with temporal information. In this paper, we propose a temporal document annotation model based on Formal Concept Analysis theory for Twitter domain. The tweets attributes defining the temporal context are the temporal expressions, the events and their types. It is also proposed a calendar especially suited to the phenomena of commemoration of anniversaries and dates in Twitter: The Social-Imaginary Calendar. The Corpus used to the experiments is a subset of the RepLab2013 collection. A detailed description of its temporal aspects is provided.
Keywords: Temporal information, Temporal annotation of tweets, Content-based information representation

1 *Introducción*

El tiempo juega un papel fundamental en todo espacio de información y Twitter[1] no es una excepción. A caballo entre red social y red de noticias, millones de personas comparten a diario, en forma de *tweet*, datos y opiniones relevantes para la reputación de personajes públicos, compañías y Gobiernos. Dos de los aspectos que mejor caracterizan Twitter son la actualidad de su contenido y el fenómeno de las *"tendencias"* (los temas más comentados en un momento dado) y ambos se miden en términos temporales.

En los últimos años Twitter ha acaparado un gran esfuerzo investigador tanto en lo que respecta a la detección de temas y tendencias como al análisis de sentimiento en los *tweets*, esto es, la polaridad de la opinión que reflejan sobre las entidades aludidas. Sin embargo, la información temporal de los *tweets*, fuera de la fecha de creación, esto es, la que se extrae del contenido, no ha recibido excesiva atención.

La explotación de esta información temporal "latente", la que se refiere a las expresiones temporales y eventos, presenta grandes desafíos y oportunidades (Alonso et al.,

[1] http://twitter.com

ISSN 1135-5948

2011)(Vicente-Díez y Martínez, 2009) y abre la puerta a la construcción de un contexto temporal complejo de los *tweets*, en el que se trata de poner en relación el momento en que son compartidos (determinado por su fecha de creación o *timestamp*) con el *momento en que sucede lo que se comparte*, para de esta manera definir qué significa la *"actualidad"* en los distintos temas que se tratan en Twitter, así como establecer eventuales relaciones de causalidad entre ellos.

Por otro lado, en un año cualquiera se dan multitud de eventos que, asociados a una fecha, permanecen en el *imaginario colectivo*, convirtiéndose en efemérides compartidas por grandes grupos de población, como es el caso del *"11 de septiembre"* a nivel mundial, del *"15 de marzo"* en España o del *"25 de junio"* para los seguidores de Michael Jackson. Con frecuencia estos eventos se reflejan en un aluvión de comentarios en Twitter, por lo que con un adecuado procesamiento de la semántica temporal, puede aprovecharse ese *"conocimiento colectivo"* para la confección de un calendario anotado con las fechas clave para una determinada entidad.

En este artículo, proponemos una aproximación a la representación de *tweets* de acuerdo a sus aspectos temporales (expresiones temporales y eventos). Las principales características de la propuesta son: utilizar el paradigma del Análisis de Conceptos Formales, como una línea de tiempo con múltiples granularidades, y articularse entorno a un calendario *dual*, basado en el *Calendario Gregoriano* y en uno definido por nosotros con el objeto de reflejar fechas señaladas, aniversarios y cualquier tipo de efeméride, el *Calendario "Imaginario-Colectivo"*.

El resto del artículo se organiza de la siguiente manera: en la Sección 2 se abordan los trabajos preliminares en anotación y representación de la información temporal asociada a un documento; en la Sección 3 se detalla la propuesta de caracterización temporal de un conjunto de documentos y se describe el desarrollo computacional para la integración y uso de herramientas y recursos de anotación, así como la explotación de estas anotaciones para la formación de los retículos temporales; en la Sección 4 se desarrollan los experimentos, se describe el corpus anotado, se extraen los descriptores y se construyen diferentes retículos en base a éstos; por último, en la Sección 5 se valoran los resultados obtenidos, se recapitula el trabajo presentado y se plantean algunas líneas de investigación abiertas que pueden ser abordadas en los próximos años.

2 Trabajos relacionados

La explotación de la información temporal contenida en un documento, cuenta con una larga tradición en la investigación, que experimentó el impulso definitivo en los años 90 con la celebración de la sexta conferencia MUC[2]. Fruto de esta investigación, que ha buscado mejorar, mediante conocimiento temporal, todo tipo de tareas en los sistemas de Recuperación de Información (detección y seguimiento de temas, búsqueda automática de respuestas, extracción automática de resúmenes, etc.) surgieron los esquemas de anotación temporal cuyo máximo exponente es TimeML (Pustejovsky et al., 2003), lenguaje que hoy es estándar. A la par que dichos esquemas, se fueron diseñando anotadores automáticos cada vez más complejos, como Tarsqi (Verhagen et al., 2005), que anota eventos y representa mediante grafos las relaciones temporales entre ellos, HeidelTime (Strötgen y Gertz, 2010), que es multilingüe y multidominio o Tipsem (Llorens, Saquete, y Navarro, 2010), también multilingüe (inglés).

Los paradigmas de visualización de la información temporal que se han utilizado van desde las líneas de tiempo y los grafos, a los mapas espacio-temporales y los grafos animados. Alonso, Gertz, y Baeza-Yates (2009) usan líneas de tiempo para agrupar resultados de búsqueda en base a las características temporales de los documentos. En un proceso similar al nuestro, extraen las expresiones temporales explícitas, implícitas y relativas de cada documento y las normalizan para crear un "perfil temporal". Este perfil se adapta a la granularidad que mejor define a la colección, según el calendario Gregoriano (Goralwalla et al., 2001) y se procede a su representacion en una línea de tiempo. Cada elemento de la línea de tiempo representará un *cluster*; obviamente puede haber *clusters* vacíos y también puede haber documentos que corresponden a más de un *cluster*, cuando tienen varias expresiones temporales; en este caso, se tratará de determinar el "*cluster principal*", atendiendo a la expresión

[2]Message Understanding Conference: www.cs.nyu.edu/cs/faculty/grishman/muc6.html

temporal predominante, esto es, que aparezca más en el documento.

En nuestra propuesta se extraen los eventos, además de las expresiones temporales, y se construye con todo ello un retículo basado en el Análisis de Conceptos Formales (en adelante, FCA, del inglés *"Formal Concept Analysis"*), que presenta dos ventajas principales: no requiere seleccionar un descriptor temporal principal, y permite el uso simultáneo de varias granularidades. En un enfoque similar, Ritter et al. (2012) extraen eventos de Twitter y los representan, de acuerdo a su contenido temporal, en un calendario que se actualiza en tiempo real[3].

Hasta donde sabemos, ningún otro trabajo ha utilizado retículos de conceptos para representación de información desde el punto de vista temporal. No obstante, el enfoque basado en FCA para modelar el contenido de conjuntos de *tweets* fue explorado por Castellanos, Cigarrán, y García-Serrano (2013) en tareas de detección de temas. Dichos autores modelan los *tweets* tomando como descriptores sus términos. Esta selección de los descriptores presenta un problema dual, por un lado hay una alta dependencia del dominio, por otro, el número de conceptos o posibles temas es potencialmente muy alto. Nuestra selección de descriptores trata de solventar estos problemas; el número de eventos anotados es sensiblemente menor que el conjunto de rasgos del vocabulario, además, las expresiones temporales son normalizables, y por tanto, independientes del dominio.

3 *Propuesta y desarrollo computacional*

3.1 Contexto temporal de una colección de documentos

Sea $\Delta = \{d1, ..., d_n\}$ una colección de documentos, su información temporal puede ser de los siguientes tipos (ver ejemplo en la Figura 1):

- **fechas de creación** o *timestamps* de los documentos,
- **expresiones temporales** contenidas en los documentos,
- **eventos** contenidos en los documentos, que se relacionan con las expresiones temporales y entre ellos.

[3]Twitter Calendar: http://ec2-54-170-89-29.eu-west-1.compute.amazonaws.com:8000/

Figura 1: Información temporal de un *tweet*

Construimos el conjunto $\tau = \Phi \bigcup T \bigcup E$ donde:

- $\Phi = \{f_1, ..., f_m\}$ es el conjunto de fechas de creación (distintas) de los documentos
- $T = \{t_1, ..., t_p\}$ es el conjunto de expresiones temporales normalizadas presentes en Δ
- $E = \{e_1, ..., e_q\}$ es el conjunto de los eventos presentes en Δ, lematizados

Se define el **contexto temporal** de Δ, $C_T := (\Delta, \tau, I)$ donde Δ es el conjunto de documentos, τ es el conjunto de atributos temporales e I es la relación binaria de incidencia que relaciona cada documento con los atributos que posee. Así construido, C_T es un **contexto formal**.

Consideramos que dos documentos pueden presentar similitud temporal, bien porque hayan sido creados en momentos temporales próximos bien porque el contenido descrito pertenezca al mismo evento o describa eventos que suceden en momentos cercanos. El retículo de conceptos formales $\beta(C_T)$ realizará un agrupamiento de los documentos que tenga en cuenta esta doble dimensión: creación/contenido.

3.2 Calendario *Imaginario-Colectivo*

La normalización de las expresiones temporales, requiere de la definición de un *"calendario"*. Lo tradicional es utilizar el calendario **Gregoriano**, que presenta las granularidades: *año, mes, día, hora, minuto* y *segundo*, con las relaciones "$\gg$" (más *gruesa*) y "$\ll$" (más *fina*) (Goralwalla et al., 2001):

$G_{\text{año}} \gg G_{mes} \gg ... \gg G_{segundo}$

La elección de una granularidad concreta no es necesaria para FCA, por el contrario, podemos representar una expresión en múltiples sistemas, con el objeto de no perder ninguna información. Por ejemplo, dadas las expresiones *"1969"*, *"1967"* y *"1967-06-01"*, si fuéramos a representar los docu-

mentos en una línea de tiempo, podríamos considerar que la granularidad más adecuada es $G_{\text{año}}$, ya que solo la última expresión se puede expresar en granularidades más finas. Sin embargo en FCA, podemos elegir el conjunto de descriptores sin perder información y a la vez mantener la relación entre dos documentos que hacen referencia al mismo año: *"1969"*, *"1967"*, *"1967-06"* y *"1967-06-01"*. El documento que posee la expresión temporal *"1967-06-01"* tendrá 3 descriptores.

Hay un tipo de expresiones que, por no estar completamente determinadas, no se pueden representar en el Calendario Gregoriano, pero tienen dimensión temporal, aunque su carácter puede ser estacional o periódico. Hablamos de expresiones del tipo *"día de Navidad"*, *"Diciembre"* o *"invierno"*, cuando **no se refieren a un año concreto**; su valor en lenguaje TimeML sería, respectivamente: "XXXX-12-25", "XXXX-12" y "XXXX-XX-XXWI". Estas expresiones no se pueden representar en una línea de tiempo al uso, sin embargo tienen cabida natural en FCA, y son relevantes en ciertos dominios, como las redes sociales, donde es frecuente hacer alusiones a todo tipo de Aniversarios o fechas señaladas.

Definimos el ***Calendario Imaginario-Colectivo*** como la terna:

$$C_{IC} = (A, \varrho, \varphi)$$

donde A representa un año natural cualquiera, $\varrho = (A_m, A_d)$ es el conjunto de granularidades (mes y día) y φ la función de conversión obvia:

$\varphi(XXXX - 12 - 25) = XXXX - 12$

Con esta definición no pretendemos capturar el significado de cada fecha para cada persona, sino ese conjunto de efemérides compartidas por un conjunto concreto de la sociedad que puede ser los seguidores de los Beatles, los habitantes de un país o la población mundial.

3.3 Desarrollo computacional

Para la construcción del retículo temporal asociado a una colección de documentos, es necesario localizar y extraer las expresiones temporales y eventos presentes en ellos y procesarlas adecuadamente para obtener el conjunto de atributos. Se ha desarrollado un entorno computacional que integra herramientas y recursos Web de anotación y FCA; su arquitectura se representa en la Figura 2 y consta de las siguientes fases:

- **Preprocesado** Se preparan los documentos para ser anotados, mediante la eliminación de los caracteres no permitidos por XML. Se eliminan también las *urls*, para evitar la anotación de fechas en las rutas de carpetas y los *emoticonos* con símbolos numéricos (como "<3"), etc.
- **Anotación** Se anotan los documentos con HeidelTime. El subconjunto de documentos en inglés se anota también con Tarsqi. No es necesario utilizar un reconocedor de idioma pues los *tweets* de RepLab están etiquetados con esta información.
- **Descriptores** Se parsean los archivos de salida de Tarsqi y HeidelTime, extrayendo, para cada documento, las fechas de creación, las expresiones temporales, los eventos y su tipo. Tras descartar las expresiones poco frecuentes o indeseadas, se enriquece el conjunto de expresiones, añadiendo su equivalencia en todas las posibles granularidades de los Calendarios definidos en la propuesta. Los eventos se lematizan usando la librería **nltk.stem.wordnet** [4]. Finalmente se genera la tabla que representa al contexto formal, constituido por los documentos (objetos) y sus descriptores temporales (atributos).
- **Retículo de conceptos** Una vez contruido el contexto formal, recurrimos al entorno de FCA Concept Explorer (Yevtushenko et al.,), para el cálculo del conjunto de conceptos formales y la representación del retículo.

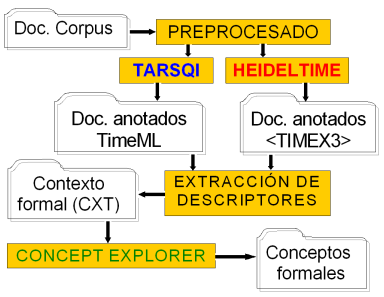

Figura 2: Diagrama funcional implementado

[4] http://www.nltk.org/

4 *Experimentación*

4.1 Anotación y estudio lingüístico del Corpus

El corpus de experimentación es un subconjunto de la colección de *tweets* de **RepLab2013**.

RepLab (Amigó et al., 2013) es un Foro de Evaluación de sistemas de gestión de reputación online. Para la edición de 2013 se seleccionaron 61 entidades de cuatro temáticas diferentes (música, universidades, banca y automóviles) y por cada una, se recogieron varias decenas de miles de *tweets*, en inglés y en español, de un periodo comprendido entre el 1 de junio y el 31 de diciembre de 2012.

Se formaron para cada entidad un conjunto de entrenamiento (unos 700 *tweets*) y uno de validación (unos 1500 *tweets*) y se anotaron los conjuntos de entrenamiento con información relativa a temática, relación con la entidad y posibles implicaciones que pudiera tener el contenido del *tweet* para la reputación de la misma. Se pretendía que los conjuntos de datos de entrenamiento y validación estuvieran formados por *tweets* distantes en el tiempo, esto es, con una brecha temporal entre ellos de varios meses. Para ello, se asignaron los primeros *tweets* al conjunto de entrenamiento y los últimos al de validación.

Para llevar a cabo la experimentación de esta propuesta, se ha elegido la entidad **"Beatles"**. El **"Corpus Beatles"**, descrito con detalle en (Vázquez-Méndez, 2014), está formado por un conjunto de entrenamiento de 701 *tweets* (538 en inglés, 163 en español) y por uno de validación de 1531 *tweets* (1130 en inglés, 401 en español).

4.1.1 Temporalizacion y temática

El periodo temporal abarcado por ambos conjuntos es bastante pequeño. El 98 % de los *tweets* de entrenamiento fueron publicados entre el 1 y el 5 de junio, mientras que el mismo porcentaje de los *tweets* de validación, lo fue entre el 22 y el 31 de Diciembre. Ubicar la colección temporalmente es fundamental, tanto para elegir la granularidad más adecuada, como para extraer conclusiones respecto a eventos cuyo periodo de vigencia coincida con el de los datos disponibles.

En cuanto a la temática, atendiendo a las anotaciones del conjunto de entrenamiento, se puede ver que los temas más habitualmente tratados están relacionados con comentarios de fans (22 %), letras y vídeos de canciones (23 %) y referencias varias a productos (ediciones remasterizadas de discos, etc.) (26 %). También se detectan varios temas que se presumen de actualidad por referirse a un evento concreto que se produce en una ventana temporal de unos pocos días respecto a la publicación del *tweet*, como son el Concierto del Jubileo (5 %) y el 45º aniversario del lanzamiento del álbum Sgt Pepper (3 %).

4.1.2 Expresiones temporales

Se anota el corpus Beatles con HeidelTime, lo que da como resultado un total de 287 *tweets*, el 16 %, con presencia de expresiones temporales (etiquetadas con TIMEX3). Se trata pues de un porcentaje pequeño pero significativo. La tipología de expresiones se resume en la Figura 3.

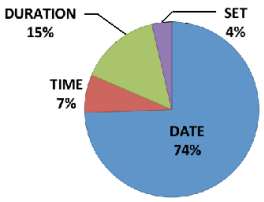

Figura 3: Tipos de TIMEX3

Predominan las expresiones de tipo "fecha" (DATE), en particular las que se refieren a tiempo presente:

- *Vamos a soñar imaginen q todos los integrantes d los beatles estuvieran vivos sería* <TIMEX3 type="DATE" value="2012-06-04">***hoy***</TIMEX3> *la locura en el concierto en honor a la reina*

Las referencias al pasado suelen aparecer de forma explícita, en *tweets* donde se data el contenido (una canción, un vídeo) o se menciona un evento de cierta relevancia, como la publicación de un álbum, un concierto, etc:

- <TIMEX3 type="DATE" value="1967-06-01">***June 1, 1967***</TIMEX3> - *The Beatles release Sgt. Pepper's Lonely Hearts Club Band. The album is certified gold its first day in stores. #TheBeatles*

El segundo tipo en importancia es "duración" (DURATION); suele tener mucho que ver con fechas señaladas en la historia de la entidad, esto es, con aniversarios o periodos en los que destaca algún aspecto de la entidad:

- *Hoy se cumplen*<TIMEX3 type="DURATION" value="P45Y"> ***45 años***</TIMEX3> *del estreno de Sgt. Pepper's Lonely Hearts Club Band, álbum de The Beatles.*

4.2 Eventos

El subcorpus de *tweets* en inglés, se anota también con Tarsqi (etiquetas TIMEX3, EVENT y LINK). Al contrario de lo que ocurría para las expresiones temporales, el porcentaje de anotación es bastante alto (62 %) y muchas veces se anotan varios eventos por *tweet* (1865 eventos anotados en 843 *tweets*). Tras el proceso de lematización, los eventos a considerar como descriptores se reducen considerablemente (646 lemas distintos).

Las palabras etiquetadas como eventos son mayoritariamente verbos, aunque también hay sustantivos o adjetivos. La influencia del dominio se hace notar en la presencia de verbos como *"listen"* o *"play"*:

- *This morning I have mostly been* <EVENT class="PERCEPTION"> ***listening***</EVENT> *to 'The Beatles - White Album'*
- *TONIGHT @thepeel *Beatles Tribute Band* Abbey Road LIVE! Sgt Pepper 45th Anniversary* <EVENT class="OCCURRENCE"> ***Show***</EVENT> *! $20! #avl #avlent #avlmusic #Asheville*
- *CHOON! The Beatles really were* <EVENT class="STATE" > ***brilliant***</EVENT> *! #jubileeconcert*

En cuanto a clase de evento, la inmensa mayoría son de ocurrencia (OCCURRENCE); también hay una presencia significativa de eventos de estado (STATE, I_STATE) y de percepción (PERCEPTION) (Figura 4).

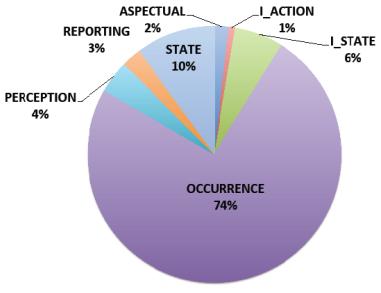

Figura 4: Clases de eventos (EVENT)

4.3 Elección de descriptores y retículos temporales

El conjunto de descriptores τ, tal como se define en la propuesta, constituye el contexto temporal más completo para un conjunto de *tweets*. Hemos realizado varios experimentos, para distintas configuraciones de τ, aumentando paulatinamente la complejidad del contexto (ver Tabla 1) y la información representada.

Experimento	**Descriptores**
I	$\tau = \Phi$
II	$\tau = \Phi \bigcup T$
III	$\tau = \Phi \bigcup T \bigcup E$
IV	$\tau = \Phi \bigcup T \bigcup clases(E)$

Tabla 1: Elección de descriptores

El retículo formado en el Experimento I (Figura 5) agrupa los *tweets* por día, mes y año, dando idea del grado de actividad de los usuarios en relación a la entidad. El tamaño de los nodos del retículo va en relación al número de objetos del concepto; por cada uno se indican los atributos (en color gris) y el número y porcentaje de *tweets* contenidos (en color blanco).

Aunque para extraer conclusiones sobre periodos de interés deberíamos contar con ventanas temporales más amplias, sí se pueden detectar alteraciones significativas. Por ejemplo, el día 4 de junio la actividad fue notablemente más alta que en el resto. La razón fue la celebración del concierto del Jubileo en honor a la Reina de Inglaterra, en el que participó Paul Mcartney y se cantaron varias canciones de los Beatles, lo que animó a los *twitteros* a comentar. Obviamente el retículo no da esta información tan concreta, pero nos alerta de que algún evento importante puede haber sucedido.

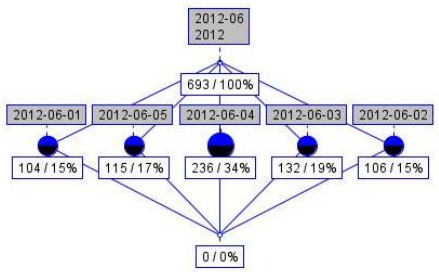

Figura 5: Diagrama de Hasse ExpI

Al introducir en el contexto las expresiones temporales presentes en el contenido (Experimento II), ya se crean agrupaciones de *tweets* que comparten ciertos atributos. Por ejemplo, el retículo ha aislado aceptablemente parte del tema *"Sgt Pepper"*. Este tema se corresponde con el nombre de un álbum de los Beatles lanzado el 1 de junio de 1967; al conmemorarse el aniversario de su lanzamiento en 2012, se convirtió en un tema comentado en Twitter. En la Figura 6 se muestra en detalle el concepto en que se basa la agrupación: prácticamente todos los *tweets* que hacen referencia al año 1967 lo hacen al álbum Sgt Pepper y todos los *tweets* escritos el 1 de junio que hacen referencia a 1967, también.

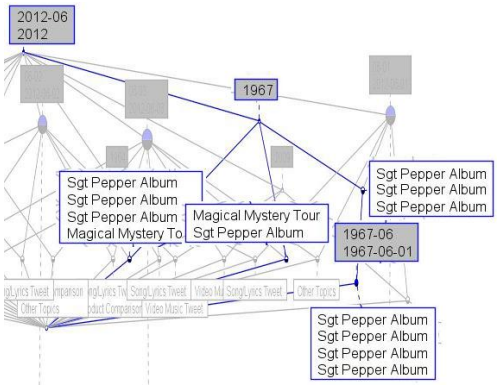

Figura 6: ExpII Tema *SgtPepper*

A continuación, agregamos las expresiones lematizadas de los eventos (Experimento III). Las reglas de asociación obtenidas por Concept Explorer permiten identificar agrupaciones de *tweets* informativas. Así se pone de manifiesto que el evento *"release"* en *tweets* con el atributo *"1967"*, se refiere exclusivamente al lanzamiento del álbum Sgt. Pepper y que todos los *tweets* escritos el 4 de junio, que contienen el evento *"sing"*, tratan del concierto del Jubileo (ver Figuras 7,8).

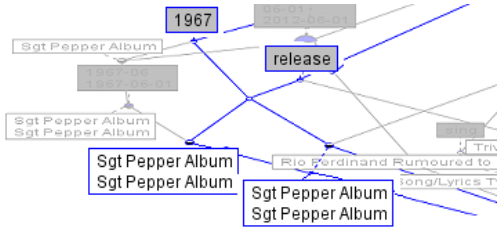

Figura 7: ExpIII Tema *SgtPepper*

Finalmente, en el Experimento IV, se decide sustituir en τ, los eventos por sus clases. De esta forma, se gana independencia respecto al dominio y se pueden detectar otro tipo de relaciones, como las que existen entre algunas clases de eventos y algunos tipos de *tweets*. Es el caso de los eventos de percepción (PERCEPTION), muy relacionados con la expresión de opiniones en distintas formas: comentarios de fans, comentarios sobre productos,etc. En la Figura 9 se muestra el retículo de conceptos para este tipo de eventos; como se ve, las percepciones sobre el concierto del Jubileo, corresponden todas al día de su celebración.

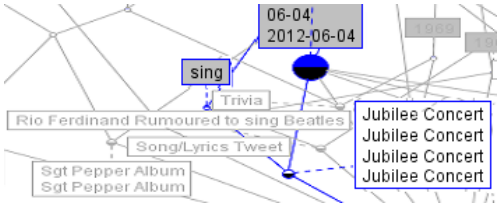

Figura 8: ExpIII Tema *Jubilee Concert*

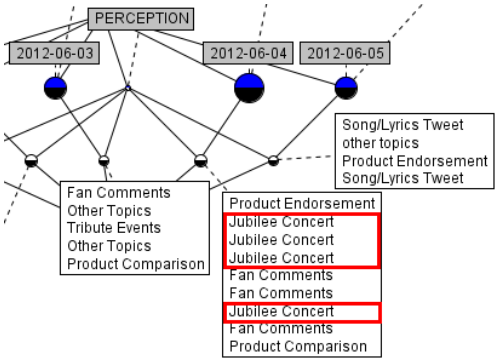

Figura 9: ExpIV Tema *Jubilee Concert*

5 Conclusiones y trabajos futuros

La explotación de la información temporal de los *tweets*, mediante su traducción en descriptores, ha permitido detectar eventos anclados a una fecha, como el aniversario del lanzamiento del album *Sgt. Pepper* y el concierto del Jubileo. Es un buen resultado, teniendo en cuenta, que el periodo de tiempo abarcado era muy reducido.

Por otro lado, la gran parte de *tweets* anotados temáticamente, lo era de un modo genérico: comentarios de fans, comentarios sobre productos, etc. Se consiguió cierto nivel de agrupación para tipos concretos de *tweets* asociados a eventos de PERCEPCIÓN; de todas formas, el modelo intentará desagregar los temas, buscando qué ha motivado cada *tweet, por qué se ha escrito en el momento en que se ha escrito*. Estas subdivisiones no

pueden ser evaluadas con las anotaciones del Corpus de las que disponemos.

La propuesta que aquí se detalla ha querido poner de manifiesto que en los *tweets*, pese a lo reducido de su extensión, hay información temporal latente que puede contribuir a mejorar el rendimiento de los sistemas en tareas como la detección y seguimiento de temas o la agrupación de documentos en Twitter.

La explotación de esta información temporal, requiere de una anotación acorde a las peculiaridades del dominio. Anotadores como Heideltime pueden configurarse para anotar textos en lenguaje *"coloquial"* para el idioma inglés, pero se necesitan Corpus de *tweets* anotados, tanto en inglés como en español, que puedan servir de *gold-standard*.

Otra línea de trabajo futuro es la anotación de *hashtags* con información temporal, como *#15m* o *#jubileeconcert*.

Bibliografía

Alonso, O., M. Gertz, y R. Baeza-Yates. 2009. Clustering and exploring search results using timeline constructions. En *Proceedings of the 18th ACM conference on Information and knowledge management*, páginas 97–106.

Alonso, O., J. Strötgen, R. Baeza-Yates, y M. Gertz. 2011. Temporal Information Retrieval: Challenges and Opportunities. *TWAW*, 11:1–8.

Amigó, E., J. Carrillo de Albornoz, I. Chugur, A. Corujo, J. Gonzalo, T. Martín-Wanton, E. Meij, M. de Rijke, y D. Spina. 2013. Overview of RepLab 2013: Evaluating Online Reputation Monitoring Systems. En *CLEF*, volumen 8138 de *LNCS*, páginas 333–352. Springer.

Castellanos, A., J. Cigarrán, y A. García-Serrano. 2013. Modelling Techniques for Twitter Contents: A Step beyond Classification based Approaches. En *Working Notes of the CLEF 2013*.

Goralwalla, I., Y. Leontiev, M.T. Özsu, D. Szafron, y C. Combi. 2001. Temporal granularity: Completing the puzzle. *Journal of Intelligent Information Systems*, 16(1):41–63.

Llorens, H., E. Saquete, y B. Navarro. 2010. TIPSem (English and Spanish): Evaluating CRFs and Semantic Roles in TempEval-2. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, SemEval '10, páginas 284–291.

Pustejovsky, J., M.J. Castaño, R. Ingria, R. Saurí, R.J. Gaizauskas, A. Setzer, G. Katz, y D.R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Ritter, A., O. Etzioni, S. Clark, y others. 2012. Open domain event extraction from Twitter. En *Proceedings of the 18th ACM SIGKDD International conference on Knowledge discovery and data mining*, páginas 1104–1112. ACM.

Strötgen, J. y M. Gertz. 2010. HeidelTime: High Quality Rule-Based Extraction and Normalization of Temporal Expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, páginas 321–324, Uppsala, Sweden, July. ACL.

Vázquez-Méndez, A. 2014. *Explotación de la Información Temporal en Twitter para la organización de tweets*. Tesis de Máster, UNED.

Verhagen, M., I. Mani, R. Saurí, R. Knippen, S.B. Jang, J. Littman, A. Rumshisky, J. Phillips, y J. Pustejovsky. 2005. Automating temporal annotation with TARSQI. En *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, páginas 81–84.

Vicente-Díez, M.T y P. Martínez. 2009. Temporal semantics extraction for improving web search. En *20th International Workshop on Database and Expert Systems Application*, páginas 69–73. IEEE.

Yevtushenko, S., J. Tane, T.B. Kaiser, S. Obiedkov, J. Hereth, y H. Reppe. ConExp - The Concept Explorer. *URL: http://conexp.sourceforge.net*.