Universidad Nacional de Educación a Distancia
Escuela Técnica Superior de Ingeniería Informática
Departamento de Lenguajes y Sistemas Informáticos

Explotación de la Información Temporal en Twitter para la organización de tweets

Tesis presentada por Asunción Vázquez Méndez
para obtener el título de Máster en Lenguajes y Sistemas Informáticos

2014

Directora: Ana García Serrano

Resumen

El tiempo es un elemento de importancia capital en todo espacio de información y Twitter no es una excepción. Antes al contrario, el nacimiento, difusión y duración de los temas tratados, las tendencias, etc. son fenómenos que se definen en términos temporales.

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.

Este Trabajo se sitúa en el campo de la Organización Temporal de la Información en Twitter. Concretamente, se propone un modelo basado en Análisis Formal de Conceptos, 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 es un subconjunto de la colección de RepLab2013, del que se hace una completa descripción en lo que concierne a sus aspectos temporales.

Finalmente, la materialización de la propuesta requiere del desarrollo de un entorno computacional para la integración y uso de herramientas y recursos de Anotación Automática y Análisis Formal de Conceptos.

Abstract

Time is a crucial element in any space of information and Twitter is no an exception. On the contrary, the creation, diffusion and duration of the topics , the trends, and other phenomena are defined in temporal terms. Although the exploitation of temporal information in Retrieval and Organization tasks has a long tradition, the approaches content-based have not been fully explored for Twitter.

This work is included in the field of Temporal Organization of Information on Twitter. Specifically, the proposal is using an approach based on Formal Concept Analysis theory. The tweets attributes defining the temporal context are the time expressions, the events and their types. It is also proposed a calender 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. The experimentation performed required the development of an environment for the computer integration and reuse of available tools and resources.

Finally, some results and future works are included.

Índice general

Índice de figuras

Índice de cuadros

Capítulo 1

Introducción

Vivimos, indiscutiblemente, en la era de la información. Nunca como hasta hoy tal cantidad de personas ha tenido tal ingente cantidad de información al alcance de la mano. El auge de Internet y los avances tecnológicos, que producen ordenadores cada más potentes y con mayor capacidad de almacenamiento, ha hecho posible esta apertura casi universal a todas las facetas del conocimiento humano. Además, la explosión de las redes sociales está transformando la forma en que las personas se relacionan, consumen, se informan, expresan su opinión, etc.

En este universo infinito, que desborda la capacidad humana, es imprescindible atender a la dimensión temporal de la información: de poco sirve saber *lo que pasa* si no sabemos ***cuándo*** *pasa*.

La investigación en la explotación de la información temporal, con el fin último de dotar a los sistemas de la capacidad de razonar temporalmente, cuenta con una larga tradición. Una de sus principales contribuciones ha sido el desarrollo de los esquemas de anotación, en base a los cuáles, la información temporal es localizada y extraida de los documentos, lista para ser utilizada en múltiples tareas de recuperación y organización de la información.

El tiempo en **Twitter**[1], como red social de noticias, es un elemento de importancia capital, con dos facetas principales: la actualidad de los contenidos y el fenómeno de las tendencias (palabras o frases más repetidas en un momento concreto). Una adecuada explotación de la información temporal que contienen los *tweets* puede

[1] https://twitter.com/

aportar beneficios al rendimiento de los sistemas que afrontan tareas de organización, detección de temas y tendencias, etc.

Este Trabajo pretende poner de manifiesto cuál es la información temporal contenida en Twitter y cómo se organizan temporalmente los *tweets*. A continuación presentamos los objetivos perseguidos, así como la metodología seguida y su estructura.

1.1. Objetivos

El objetivo principal de este Trabajo es la puesta en valor de la información temporal asociada a los *tweets*, para su explotación en diferentes tareas de recuperación y organización. Se pretende construir un modelo de representación temporal de los *tweets* basado en Análisis Formal de Conceptos, una teoría matemática de reciente aplicación a tareas de agrupación de *tweets* basadas en contenido.

La utilización de la información temporal asociada a un documento requiere de su definición, extracción y normalización. En concreto, se plantean las siguientes cuestiones, a las que trataremos de responder:

- *¿Cuál es la información temporal asociada a un documento?* Veremos que hay, *grosso modo*, dos tipos de información temporal, una asociada a los metadatos (fecha de creación, modificación, etc) y otra incrustada en el contenido, la información temporal "latente".

- *¿Cómo se expresa el tiempo en Lenguaje Natural?* Reflexionaremos sobre qué es el tiempo y cuál es su lógica y qué palabras o expresiones sirven de vehículo a la información temporal en el lenguaje (fechas, adverbios, verbos,...).

- *¿Cómo localizar automáticamente la información temporal contenida en un texto?* Haremos un repaso por las distintas aproximaciones, en general, basadas en reglas manuales y muy dependientes del idioma.

- *¿Cómo traducir las expresiones temporales y normalizarlas, de forma que tengan realmente un valor temporal?* En particular, *¿cómo hacerlo si las expresiones son ambiguas o relativas?* Veremos cómo aprovechar el contexto de

la expresión, tanto en lo relativo a la creación del documento, etc. como a cualquier otra referencia que aparezca en el contenido.

- *¿Qué se entiende por "evento"? ¿Cómo se pueden ordenar temporalmente dos eventos y cuál es su utilidad?* En nuestro contexto, un "evento" será "algo que sucede en un momento dado". Veremos que la identificación de eventos y su seguimiento es un aspecto clave en distintas tareas de organización de la información.

Por otro lado, la aplicación del Análisis Formal de Conceptos a la representación del contenido de un corpus de documentos, requiere de la definición de un conjunto de atributos o descriptores que los caractericen. La Propuesta que se planteará, responderá a lo siguiente:

- *¿Cuáles son los descriptores temporales que mejor describen a un conjunto de tweets?*

1.2. Metodología

La metodología de trabajo propuesta para la consecución del objetivo principal, así como para dar respuesta a las cuestiones planteadas es la siguiente:

1. Estudio del estado del arte en representación y anotación de la información temporal, y en herramientas y recursos de anotación automática.
2. Estudio de la teoría matemática del Análisis Formal de Conceptos (AFC) y de su aplicación a la organización de resultados de búsqueda.
3. Estudio del estado del arte en explotación de la información temporal en distintas tareas de la organización de información.
4. Estudio y propuesta de un conjunto de descriptores de índole temporal para la organización de información mediante AFC.
5. Desarrollo del entorno computacional para el procesado y anotado temporal de un conjunto de *tweets* y la extracción de sus descriptores temporales.

6. Estudio de las características temporales del corpus y desarrollo de diferentes experimentos, basados en los descriptores propuestos, para la agrupación de *tweets*. Valoración de resultados.

1.3. Estructura del Trabajo

El trabajo está estructurado en tres partes:

- *Estado del arte*. Se abordan los trabajos preliminares en anotación y explotación de la información temporal asociada a un documento y en aplicación del Análisis Formal de Conceptos a la organización de información basada en contenido.
- *Propuesta y experimentación*. Se presenta el modelo de aproximación mediante Análisis Formal de Conceptos a la organización de información basado en el contenido temporal. Se describe el desarrollo computacional para la integración y uso de herramientas y recursos de anotación. Se desarrollan los experimentos y se valoran los resultados.
- *Conclusiones finales*. Se recapitula el trabajo presentado y se plantean algunas líneas de investigación abiertas que pueden ser abordadas en los próximos años.

Parte I

Estado del Arte

Capítulo 2

La Información Temporal

En este capítulo se introducirá brevemente el contexto en que, a principios de los años 80, cobra fuerza el interés en dotar de razonamiento temporal a los sistemas informáticos. Presentaremos, en líneas generales, la lógica temporal de Allen, en la que se basan la mayor parte de los sistemas de representación del tiempo y nos detendremos en los conceptos clave de *intervalo*, *calendario* y *granularidad*. Expondremos también, de modo exhaustivo cómo se expresa el tiempo en Lenguaje Natural. A continuación veremos cómo surge la tarea de Anotación Temporal y cuál ha sido su evolución. Presentaremos los primeros esquemas de anotación, y también las completas guías diseñadas para la construcción de corpus anotados con información temporal, como *TimeBank*. Describiremos en detalle *TimeML*, el lenguaje estándar para la anotación temporal y haremos un repaso de los principales sistemas de anotación automática.

2.1. El lenguaje del tiempo

Para un ser humano, el paso del tiempo es algo natural. Vivimos el presente siendo plenamente conscientes de la realidad del pasado, del carácter potencial del futuro, de las relaciones de causalidad entre gran cantidad de eventos y en general, del cambio. Ser conscientes del cambio y la rapidez con que reaccionamos ante él, modificando nuestros esquemas mentales, es lo que nos hace inteligentes.

Paradójicamente, en el campo de la Inteligencia Artificial, muchas veces se han

dejado de lado consideraciones temporales para *evitar complicaciones*. Así, como expone **McDermott** en *"A temporal logic for reasoning about processes and plans"* [30]:

"Because time has been neglected, medical diagnosis programs cannot talk about the course of a disease. Story understanding programs have trouble with past events. Problem solvers have had only the crudest models of the future, in spite of the obvious importance of future events."

Motivado por estos problemas, el autor, construye en dicho trabajo una lógica temporal robusta con el objeto de que sirva de marco a aquellos programas que deban manejar cuestiones temporales.

En la misma época, en el área del Procesamiento del Lenguaje Natural, **Allen** observa que los esfuerzos investigadores han estado dirigidos sobre todo a la extracción y captura de la información temporal contenida en un discurso; pero este conocimiento ha de ser capaz de responder a cuestiones relacionadas con el discurso en un momento posterior y para ello hace falta una forma más potente de representación del conocimiento temporal. Con este objeto, Allen construye una lógica temporal basada en intervalos [2], que sigue siendo, hoy en día, referencia obligada para todos aquellos investigadores que abordan la dimensión temporal de la información.

Ambos trabajos tiene una característica común, y es que ninguno abandona la perspectiva computacional, sino que presentan sendos algoritmos de razonamiento temporal. La diferencia principal entre ellos es que, mientras la lógica de Mcdermott se basa en puntos o instantes temporales, Allen escoge identificar los momentos con intervalos.

2.1.1. La lógica temporal de Allen

Allen parte de la premisa de que todo modelo de representación temporal debería ser capaz de permitir la imprecisión, la incertidumbre, operar con distintas granularidades y soportar cierto grado de persistencia [2]. Veamos a qué se refieren estos conceptos en nuestro contexto.

La *imprecisión* está relacionada con el conocimiento temporal estrictamente relativo: se sabe que A es anterior a B, pero se ignora cuánto anterior, cuándo pasa A o cuándo pasa B. La *incertidumbre* aporta aún un grado mayor de relatividad, pues

la relación entre A y B no es explícita, sino que se conocen ciertas restricciones sobre el modo en que podrían relacionarse. La *granularidad* es la medida de tiempo (años, días, horas, ...) que elegimos para representar un conocimiento; depende obviamente del dominio al que pertenezca la información, puede ser de años para el conocimiento histórico o de nanosegundos para el diseño de un programa informático. Por último, la *persistencia* tiene que ver con cómo se mantiene el estado de las cosas cuando no han cambiado. En el ejemplo de Allen, se identifica con razonamientos del tipo *"Si aparqué mi coche junto al parque el lunes, y no he vuelto a moverlo, debería de estar allí ahora"* (obviamente la certeza no existe, pues pueden habérmelo robado).

Puntos vs intervalos

Allen basa su modelo, que busca responder a todas las cuestiones anteriores, en la noción de intervalo. Hace esta elección, en detrimento de los puntos, apoyándose en que, en general, en el lenguaje (en concreto en el idioma inglés) las referencias a relaciones temporales son vagas e implícitas. Por otro lado, incluso cuando parece que un evento se produce en un instante de tiempo, es probable que podamos dividirlo en sub-eventos, dándonos cuenta de que sólo era un acontecimiento instantáneo en apariencia. Veámoslo con ejemplos:

(1) *We found the letter at twelve noon*

(2) *We found the letter while John was away*

En el ejemplo (1) parece que el evento de encontrar la carta se produjo en un momento preciso del tiempo mientras que en (2) la conjunción temporal "while" indica que el momento en que apareció la carta ocurrió durante el tiempo que John estuvo fuera. Una representación basada en intervalos parece responder mejor a los dos ejemplos que una basada en puntos. Más aún, el evento "encontrar la carta", que en apariencia es instantáneo puede ser divido en sub-eventos como: mirar hacia el lugar en el que está la carta, visualizar un objeto y asociar mentalmente que dicho objeto es la carta; obviamente aunque estos sub-eventos son sucesivos y muy rápidos, no se producen de modo simultáneo.

Relaciones temporales entre intervalos

Pero el aspecto fundamental de la teoría de Allen y el que más aplicaciones ha tenido, es su definición de las relaciones posibles entre dos intervalos.

Un **intervalo temporal** se define como el momento en que sucede un evento.

Dados dos intervalos temporales, X e Y, momentos en que suceden los eventos E y E', ha de darse alguna de las siguientes situaciones:

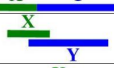
1. **X** $<$ **Y** : el evento E ocurre antes que el evento E'
2. **X** $=$ **Y**: el evento E y el evento E' ocurren simultáneamente
3. **X (se solapa con)** ***overlaps*** **Y**: el evento E termina cuando el evento E' ya ha comenzado
4. **X (se extiende hasta)** ***meets*** **Y**: el evento E termina justo cuando comienza el evento E'
5. **X (durante)** ***during*** **Y**: el evento E sucede mientras el evento E' está sucediendo

La relación *during* puede subdividirse en tres sub-relaciones: X (durante) *during* Y; X (comienza cuando) *starts* Y y X (acaba cuando) *finishes* Y.

Considerando éstas nuevas relaciones y las relaciones inversas de todas ellas, existen trece posibles relaciones entre dos intervalos, que se resumen en el Cuadro 2.1.

Paralelamente se construye una **tabla de transitividad** en la que dados tres intervalos X, Y, Z, de forma que $Xr1Y$ e $Yr2Z$, siendo $r1, r2$ alguna de las doce relaciones antes definidas (excluida la igualdad) se deducen las posibles relaciones entre X y Z. Por ejemplo, sean X, Y, Z tales que XmY, YfZ, ¿qué sabemos sobre la relación entre X y Z? Gráficamente (Figura 2.1) puede verse que las posibilidades son: XdZ, XoZ ó XsZ.

Figura 2.1: Ejemplo de transitividad

Relación	Símbolo	Símbolo inversa	Representación
X before Y	<	>	
X equal Y	=	=	
X meets Y	m	mi	
X overlaps Y	o	oi	
X during Y	d	di	
X starts Y	s	si	
X finishes Y	f	fi	

Cuadro 2.1: Las trece relaciones de orden entre intervalos

Con estos elementos, Allen construye un algoritmo en el que el conocimiento temporal se estructura en una **red**; cada **nodo** es un intervalo y la relación con otros intervalos se explicita etiquetando el **arco** que los une. Cuando dos nodos no están unidos directamente pero existe un camino que los une, pasando por otros nodos, la tabla de transitividad nos dará la relación o relaciones posibles entre ellos.

2.1.2. Calendarios y granularidades

La información temporal, tanto la que se refiere a momentos temporales, como fechas y horas, como la que expresa duraciones, puede manifestarse en distintas unidades: meses, días, años, minutos,... a las que se denomina, genéricamente, *granularidades*.

Como dijimos antes, un sistema con capacidad de razonamiento temporal ha de ser capaz de convertir unas granularidades en otras para operar con ellas. **Goralwalla** et al. [25] formalizan la representación de la información temporal a través de la definición de un *calendario*.

Según se expone en dicho trabajo, un **calendario** es una terna

$$C = (G_{tl}, \varrho, \varphi)$$

donde G_{tl} es la línea de tiempo, $\varrho = (G_1, ..., G_n)$ es el conjunto de granularidades y φ las funciones de conversión entre éstas.

Cada *granularidad* está asociada a una unidad de medida del tiempo (años, meses, días, ...) y una granularidad será *más fina* que otra si la unidad temporal en que se basa es más pequeña, esto es, "dura menos" que la otra.

En el **calendario Gregoriano**, por ejemplo, las granularidades serían:

$\varrho = (G_{\text{año}}, G_{mes}, G_{\text{día}}, G_{hora}, G_{minuto}, G_{segundo})$

con las relaciones "$\gg$" (más *gruesa*) y "$\ll$" (más *fina*) :

$G_{\text{año}} \gg G_{mes} \gg G_{\text{día}} \gg G_{hora} \gg G_{minuto} \gg G_{segundo}$.

$G_{\text{año}}$ es más gruesa que G_{mes}, G_{mes} más gruesa que $G_{\text{día}}, ...$ y por tanto $G_{\text{año}}$ es más gruesa que $G_{\text{día}}$, etc.

Las *funciones de conversión*, φ, darían el número de unidades de una granularidad que contiene otra más gruesa. Por ejemplo:

$f_C^{\text{año}}(1995) = 12$ esto es, el año 1995 tuvo 12 meses

$f_C^{\text{mes}}(1995, 9) = 30$ esto es, el mes de Septiembre del año 1995 tuvo 30 días

2.1.3. La estandarización del tiempo: la norma ISO 8601

La representación de la información temporal presenta un alto grado de ambigüedad: en unos países la semana empieza el lunes mientras que en otros lo hace en domingo; el 4/5/2014 es el 4 de mayo en España, pero el 5 de abril en el mundo anglosajón; las 4:00 pueden ser las 4 de la mañana o las 4 de la tarde; hay diferentes zonas horarias, horarios de invierno, etc.

Como consecuencia de esto, el intercambio de información temporal entre distintos sistemas no será posible a no ser que ambos conozcan el formato de representación que usa el otro. Para resolver este tipo de cuestiones se utilizan los "estándares".

El principal estándar de tiempo es la norma **ISO 8601:2004**[1] [54], que sigue el criterio de especificar en primer lugar los periodos de tiempo más largos. Así, la notación "2014-05-04"sería la estándar para el ejemplo anterior.

Principios generales

Para evitar ambigüedades cada valor (año, mes, día, hora del día) tiene un número fijo de dígitos que debe ser completado con ceros. Se prefiere, además, el sistema de 24

[1] http://www.iso.org/iso/home/store/catalogue_tc/catalogue_detail.htm?csnumber=40874

horas frente al de 12. De esta forma, el orden cronológico coincide con el lexicográfico.

Existen dos formatos, el **básico**, con el mínimo número de caracteres y el **extendido**, con separadores para aumentar la legibilidad. Se utiliza un guión (-) como separador entre los elementos de la fecha y dos puntos (:) como separador entre horas, minutos y segundos. Volviendo al ejemplo, se escribirá "20140504" en formato básico y "2014-05-04" en formato extendido.

La representación completa ha de incorporar todos los elementos de fecha u hora, pero existe una representación de precisión reducida para los casos en que se ignore alguno de estos elementos; siempre que los que se conozcan sean los más representativos. Por ejemplo "2014-05" es una fecha ISO 8601 válida, que indica el quinto mes del año 2014 (nunca el día 5 de algún mes del año 2014).

Representaciones

Fechas

Como dijimos antes, han de usarse siempre 4 cifras para el año, dos para el mes y dos para el día del mes, rellenando con ceros si es necesario. La **representación completa** permite especificar un día concreto del calendario, mientras que la de **precisión reducida** permite especificar un mes, un año o un siglo concretos (ver Cuadro 2.2.

Formato básico	Formato extendido
YYYYMMDD (*20140504*)	YYYY-MM-DD (*2014-05-04*)
YYYYMM (*201405*)	YYYY-MM (*2014-05*)
YYYY (*2014*)	-
YY (*20*)	-

Cuadro 2.2: Fecha - representación completa y reducida

Se puede también representar la llamada **fecha ordinal** (número de días transcurridos en el año), así como las distintas semanas del año, que han de especificarse con dos dígitos y precedidas del carácter "W". La primera del año se considera aquélla que contiene el primer jueves del año, o, equivalentemente, el día 4 de enero. Asimismo, los días de la semana se representan numéricamente con un dígito, comenzando en lunes (el lunes es el día 1 y el domingo el 7, por tanto) (ver Cuadro 2.3).

	Formato básico	Formato extendido
Fecha ordinal	YYYYDDD (*2014124*)	YYYY-DDD (*2014-124*)
Fecha semanal	YYYYWwwD (*2014W187*) YYYYWww (*2014W18*)	YYYY-Www-D (*2014-W18-7*) YYYY-Www (*2014-W18*)

Cuadro 2.3: Fecha ordinal y fecha semanal

Horas

Se usan dos dígitos para la hora (entre 00 y 24), dos para los minutos (entre 00 y 59) y dos para los segundos (entre 00 y 59). Se antepone el señalador "T" para marcar que se refiere a una hora local; para la hora en el sistema UTC, se añade el señalador "Z". Por ultimo, la hora local relativa a UTC (*Universal Time Coordinated*), se expresa añadiendo a la expresión de la hora, la diferencia horaria (signo "+" o "-"seguido de las horas u horas y minutos de diferencia).

Fecha y hora se pueden expresar conjuntamente utilizando el separador "T" entre ambas. Por ejemplo: 2014-05-04T18:21.

Finalmente, la **duración** se representa siempre precedida del señalador ("P") y con un señalador específico para cada componente de tiempo.

2.1.4. El tiempo en el Lenguaje Natural

La información temporal se expresa, en el lenguaje natural, de formas diversas.

Pensemos en los siguientes ejemplos:

(1) *Llegaré* ***el viernes a las 22 horas***

(2) *Llegaré* ***después del partido***

Tanto (1) como (2) tienen carácter temporal. La primera se refiere a un momento concreto que podemos fijar en un calendario; la segunda, por el contrario es menos precisa y sin embargo, también tiene un claro valor temporal.

Algunos autores, como Allen [3] y Galton [21] han trabajado en la definición de una *ontología temporal*; en otros trabajos, de corte más pragmático, como el de Schilder el al. [43] se definen dos tipos básicos de expresiones temporales, las que denotan tiempo y las que denotan eventos. En esta línea, pero con mayor vocación de generalidad, se sitúa el trabajo de Setzer y Gaizauskas [46], que tomaremos de referencia por su influencia en los esquemas de anotación actuales. Para éstos, la in-

formación temporal puede ser de los siguientes tipos: **eventos, estados, momentos** y **relaciones temporales**.

Eventos

Un *evento* es algo que ocurre en un momento, algo que podemos situar en una línea de tiempo. Los eventos suelen expresarse mediante un verbo (3) o una nominalización (4):

(3) *El partido se **jugó** bajo la lluvia.*

(4) *El partido sigue **suspendido** por la lluvia.*

Los eventos pueden clasificarse en:

- De ocurrencia (*ocurrence events*): son los más comunes, los que describen las cosas que pasan, como en (3).
- De narración (*reporting events*): asocian un evento (de ocurrencia) con la fuente de la información. Veamos un ejemplo:

  (5) *El arbitro **dice** que el partido se suspende.*

- De percepción (*perception events*): se refieren a la percepción física de otro evento:

  (6) ***He oído** pasar el camión de la basura.*

- Intencionales (*attitude events*): expresan la voluntad de llevar a cabo un evento (7) o la creencia de que ocurrirá (8). No garantizan que el evento llegue a producirse (7) o se haya producido (8):

  (7) ***Intentaré** llegar antes de la hora de la cena.*

  (8) ***Creo** que he perdido el tren.*

- Aspectuales (*aspectual events*): tienen que ver con cierta parte en la estructura temporal de un evento (el principio, el desarrollo o el final):

  (9) *En diez minutos, **empiezo** a hacer la pizza.*

Hay conjuntos de verbos que son característicos de un tipo de evento. Por ejemplo, verbos como *escuchar, notar, ver, vislumbrar...* se asocian con eventos de percepción; verbos como *decir, narrar, contar, exponer, manifestar...* lo hacen con eventos de narración, etc.

Estados

Un *estado* es como una *"fotografía"* de la realidad en un momento dado, es una situación en la que cierta entidad mantiene un atributo con ciertas propiedades o en la que varias entidades se relacionan de cierta manera. Pensemos en el ejemplo:

(10) *Tengo 35 años.*

No se puede decir que haya sucedido nada, por lo que no es un evento, sino una situación que se mantiene durante cierto tiempo.

Lo que constituye un evento es un **cambio de estado**. Volviendo al ejemplo, cuando tenga 36 años, habrá sucedido un evento: "*cumplir años*".

Los estados pueden situarse en un línea temporal y son muy importantes en lo que respecta a establecer un contexto temporal.

Momentos

Los *momentos* son los objetos temporales más intuitivos; aquéllos que se pueden situar exactamente en un calendario o en un planning horario, dependiendo de su granularidad. Algunos ejemplos son:

(11) *Llegó a las* ***12h****.*

(12) *Habrá que esperar al* ***segundo cuatrimestre****.*

(13) ***El próximo viernes*** *estrenan la película.*

(14) *La Edad Media comienza en* ***1453****.*

(15) *Su cumpleaños es* ***el 22 de febrero****.*

(16) ***Hoy*** *no me levanto.*

(17) *La exposición se inauguró* ***el Día del Libro****.*

Fechas y *horas* pueden expresarse de distintas formas numéricas (11 y 14) y mediante los nombres de los meses del años (15), de los días de la semana (13), de las partes del día (*mañana, tarde, noche, mediodía*), de las distintas divisiones del año (*bimestres, trimestres, cuatrimestres, semestres*) (12) o de las agrupaciones de varios

años (*bienio, trienio, quinquenio, década, siglo, milenio*). También hay periodos o momentos que por cualquier razón reciben un nombre, como *Semana Santa* o el *Día de San Valentín* (17). Están además las expresiones relativas al presente: *hoy, ayer, mañana* (16) y aquéllas que, mediante modificadores, utilizan las distintas granularidades para situar un momento respecto al actual:

(18) ***Hace tres días*** *que no veo al gato.*

(19) *Me encantaron las vacaciones* ***del año pasado****.*

(20) *Apareció por aquí* ***varias horas después****.*

Las expresiones temporales que denotan momentos suelen clasificarse en tres tipos [6, 43]:

- **Explícitas**: se corresponden directamente con una fecha o un momento del día (11, 14, 15),
- **Relativas**: necesitan una **referencia temporal** para ser situadas exactamente en una línea de tiempo (13, 16, 18, 19). Dicha referencia puede ser el tiempo presente, en un discurso hablado o el día en que se publicó una noticia, un tweet, etc,
- **Implícitas**: Se refieren a periodos o eventos que reciben un nombre, como en (17); se pueden situar en una línea de tiempo siempre que se conozca la correspondencia entre la expresión y la fecha. En dicho ejemplo, podemos situar la inauguración de la exposición el 23 de abril (*Día del Libro*).

Existen otro tipo de expresiones temporales, denominadas **"vagas"**, que no pueden ser situadas exactamente en una línea de tiempo aunque aportan información temporal, como en (20).

Relaciones temporales

Los eventos pueden estar relacionados temporalmente unos con otros; éstas relaciones suelen identificarse en el discurso con conectores temporales como preposiciones, conjunciones, etc.:

(21) *Me levanté* ***antes*** *de que saliera el sol.*

(22) ***En cuanto*** *vengas, salimos.*

Los tipos de relación (*antes, después, mientras,...*) son equivalentes al conjunto de relaciones entre intervalos presentado por Allen en su teoría y que vimos en el Apartado 2.1.1, por lo que no nos extenderemos.

2.2. Anotación temporal

El aprovechamiento de la información temporal contenida en un documento pasa por la localización y situación temporal de eventos y expresiones temporales, así como por el establecimiento de relaciones de orden entre ellos. Los **esquemas de anotación** surgen tanto para asistir en estas tareas al anotador humano, como para el diseño de sistemas automáticos.

La anotación temporal es una sub-tarea del Reconocimiento de Entidades Nombradas, con veinte años de tradición en los foros de evaluación competitiva. La investigación en la materia se ha apoyado en corpus como **TimeBank**, anotados manualmente con información temporal.

A pesar de los avances que se han producido en anotación automática, el problema del reconocimiento y contextualización de eventos sigue abierto, teniendo un gran impacto en el rendimiento de los sistemas tanto el idioma como el dominio al que pertenecen los documentos.

2.2.1. Esquemas de anotación

En el apartado anterior vimos que la información temporal se compone de eventos y expresiones temporales, y de las relaciones entre ellos. Los esquemas de anotación dan pautas para el **marcado** y **normalización**, de estas entidades y para hacer explícitas las relaciones temporales entre ellas. El **marcado** se refiere a la identificación del lugar exacto que ocupa la expresión en el texto, qué palabras la forman, y suele hacerse encerrando la expresión entre etiquetas SGML o XML. La **normalización** busca darle un formato estándar, que respete su semántica, independientemente del idioma o los términos utilizados; en el caso de expresiones no explícitas, es una tarea compleja que requiere del conocimiento del contexto de la expresión, así como de una referencia, que puede ser la fecha de creación (***timestamp***) u otra presente

en el documento. La determinación de la referencia más adecuada ha sido abordada por varios autores [20, 29, 38]. En ocasiones, conocer el instante de referencia es insuficiente para resolver la expresión, porque ésta es vaga o indeterminada. P.ej *"el lunes"* puede referirse al próximo o al pasado; en estos casos puede ser de utilidad atender al tiempo verbal del evento. Otra veces es necesaria incluso la aplicación de una función temporal (conocer el calendario), como por ejemplo en la expresión *"el Día de la Madre"*, que es el primer domingo de Mayo y por tanto, un día diferente cada año.

En los primeros tiempos, cada grupo de investigación diseñaba su propio esquema de anotación. Por ejemplo, **Mani** y **Wilson**, en su trabajo *"Robust temporal processing of news"* [29], se apoyaban en la etiqueta TIMEX que podía ser de tipo TIME o DATE y que representaba las expresiones como puntos de una línea temporal, con un valor (VAL) en el estándar ISO 8601 (en su versión de 1997). Su mérito es, que al contrario que trabajos anteriores, afrontaban la resolución de expresiones relativas; también hicieron un primer intento de construcción de una cronología de eventos. Una de las principales carencias de su esquema es que analizaban las expresiones sin tener en cuenta las preposiciones que las acompañaban (esto es, *"el viernes"*, *"antes del viernes"* o *"después del viernes"* serían etiquetadas igual). El trabajo de **Schilder** y **Habel** [43], *"From temporal expressions to temporal information: semantic tagging of news messages"*, supera al trabajo anterior en este aspecto, considerando preposiciones temporales mediante las que anotan siete tipos de relaciones (*before, after, incl., at, starts, finishes, excl.*). Ya mencionado en el capítulo anterior, **Setzer** y **Gaizauskas** [46] definen un esquema mucho más complejo, que anota eventos (EVENT) prestando atención a su tipo (*ocurrence, percecption, etc.*) y expresiones temporales (TIMEX) así como las relaciones temporales entre ellos mediante la etiqueta SIGNAL.

Todos los trabajos mencionados tienen en común haberse desarrollado para el procesamiento de textos de noticias. Respecto al idioma, salvo el de Schilder y Habel, que etiquetaba textos en alemán, están diseñados para el inglés. Con los avances en la investigación y el creciente interés en el procesamiento temporal, se vio la necesidad de construir un lenguaje estándar para la anotación. Así, surgieron las guías de anotación *TIDES* [18, 19] y el lenguaje, que hoy es estándar ISO, *TimeML* [34].

Con respecto al **español**, destacan las aportaciones de **Saquete** y **Martínez-Barco** [38, 39], coetáneas de los trabajos antes mencionados. Más adelante, se asumieron los estándares TIDES y TimeML, aunque teniendo en cuenta las particularidades del español en la formulación de las reglas manuales para la detección de las expresiones temporales [52]. Finalmente, con motivo de la celebración del Congreso TempEval-2010 (del que trataremos en el próximo apartado), se elaboraron dos completos informes técnicos, guías de anotación tanto para expresiones temporales [42] como para eventos [41]. Dichas guías recorren exhaustivamente las peculiaridades del español y sus diferencias con el inglés (verbos auxiliares, modos, ...), definiendo nuevos atributos cuando es necesario y estableciendo normas precisas para la anotación de expresiones complejas.

TIDES

El estándar **TIDES** (denominado así por haber nacido bajo el auspicio de la DARPA[2], en su programa *Translingual Information Detection, Extraction and Summarization*), se basa en la etiqueta TIMEX2, con los atributos resumidos en el Cuadro 2.4.

Atributo	**Función**	**Ejemplo**
VAL	Expresión normalizada de fecha u hora	VAL=“1964-10-16”
MOD	Modificador temporal	MOD=“APPROX”
ANCHOR_VAL	Expresión normalizada de una referencia temporal	ANCHOR_VAL=“1964-10-16”
ANCHOR_DIR	Relación entre ANCHOR_VAL y VAL	ANCHOR_DIR=“BEFORE”
SET	Expresiones que denotan periodicidades	SET=“YES”
NON_SPECIFIC	Identificador de expresiones indeterminadas	NON_SPECIFIC=“YES”
COMMENT	Comentarios del anotador	COMMENT=“contexto ilegible”

Cuadro 2.4: Atributos de TIMEX2 [18]

[2]Defense Advanced Research Projects Agency: http://www.darpa.mil

Su completa guía de anotación [18, 19] detalla qué tipo de expresiones se anotan, cuáles son las palabras clave, cómo se captura su significado y cuál es su extensión.

Un ejemplo de texto anotado sería [18]:

< TIMEX2 VAL="2014-05-09NI" MOD="EARLY">**early last night**</TIMEX2>

TIDES sirvió de base para la creación de un corpus multilingüe, manualmente anotado, procedente de diálogos en inglés y español, denominado *TIDES Temporal Corpus*. En 2005 se publicó la última versión del estándar, que perdió relevancia con el nacimiento y consolidación de TimeML, por el que fue absorbido.

TimeML

TimeML es un lenguaje basado en XML para la anotación de eventos y expresiones temporales. Fue creado en el ámbito del congreso **TERQAS** (*Time and Event Recognition for Question Answering Systems*) en 2002.

TimeML tomó como base el estándar de anotación TIDES [19] y el lenguaje de anotación temporal presentado por **Setzer** en su tesis [45].

Al contrario que la mayor parte de trabajos anteriores, TimeML separa la representación de eventos y expresiones temporales de su situación temporal y del orden relativo entre distintos eventos localizados en un texto. Para ello define cuatro estructuras de datos, cuatro etiquetas: EVENT, TIMEX3, SIGNAL y LINK.

EVENT

Hay varios tipos de eventos, según la naturaleza de la acción descrita (su grado de culminación, etc.). En el Cuadro 2.5 se exponen los tipos de eventos que se contemplan en la última versión de TimeML, la 1.2.1.

Ejemplo: *John left two days before yesterday*

El verbo "left" es anotado como evento, de la siguiente manera:

John <EVENT eid="e1" class="OCCURRENCE">**left**</EVENT> two days before yesterday

TIMEX3

La etiqueta TIMEX3 se usa, análogamente a TIMEX2, para marcar expresiones temporales. De hecho los criterios sobre qué elementos se marcan, así como la exten-

Atributo	**Descripción**
EventID	Identificador único de cada evento
Class	REPORTING - Describen la acción de una persona u organización declarando algo, narrando un evento, etc
	PERCEPTION - Llevan implícita la percepción física de otro evento
	ASPECTUAL - Refieren un momento dentro de la historia del evento (inicio, fin, continuación,...)
	I_ACTION - Llevan asociado un argumento del que se infiere algo, según su relación con el evento en sí
	I_STATE - Llevan asociado un estado de las cosas (posible o alternativo) del que se infiere algo, segun su relación con el evento
	STATE - Describe circunstancias en las que algo ocurre o se manifiesta cierto
	OCCURRENCE - Describe el resto de eventos que describen algo que pasa u ocurre en el mundo
Stem	(Opcional) Lexema del núcleo del evento

Cuadro 2.5: Atributos de EVENT

sión de lo marcado y el uso de una o varias etiquetas, coinciden, en general con lo establecido en la guía de anotación de TIDES [19]. En la guía de TimeML se dedica un apartado a explicar las diferencias entre ambos esquemas [36].

En el cuadro 2.6 describimos los principales atributos de la etiqueta TIMEX3.

Ejemplo: *John left two days before yesterday*

Supongamos que la fecha de creación del documento, que ha de servir de referencia es t0="2014-05-09". La expresión "two days before yesterday"(t3) se descompone en dos: "two days" (t1) del tipo duración y "yesterday" (t2) de tipo fecha.

John left <TIMEX3 tid="t1" type="DURATION" value="P2D" beginPoint="t2" endPoint="t3">**two days**</TIMEX3> before <TIMEX3 tid="t2" type="'DATE" value="2014-05-08" temporalFunction="true" anchorTimeID="t0">**yesterday**

</TIMEX3>

<TIMEX3 tid="t3" type="DATE" value="2014-05-06" temporalFunction="true" anchorTimeID="t1"/>

Atributo	**Descripción y valores**
Timex ID number	Identificador único de cada expresión temporal
Type	DATE - Expresiones que se refieren a fechas
	TIME - Expresiones que denotan partes del día
	DURATION - Expresiones que definen intervalos de tiempo
	SET - Expresiones temporales compuestas, que denotan frecuencias, etc.
Value	Contiene una expresión normalizada de la fecha/hora representada o de la granularidad, en el caso de las duraciones. Por ejemplo *"XXXX-10-20"* para *Oct 20th* o *"P4D"* para *four days*
Mod	(Opcional) Captura la semántica asociada a los modificadores de la expresión temporal, mediante los valores:= BEFORE, AFTER, ON_OR_BEFORE, ON_OR_AFTER, LESS_THAN, MORE_THAN, EQUAL_OR_LESS, EQUAL_OR_MORE, START, MID, END y APPROX
temporalFunction	(Opcional) Indica si el valor de la expresión ha de ser calculado mediante la aplicación de una función temporal
anchorTimeID	(Opcional) Indica el ID de la expresión temporal que sirve de referencia a la expresión que se está anotando
valueFromFunction	Se usa en la anotación automática
functionInDocument	Constituye una referencia para otras expresiones temporales del documento. Sus valores posibles son: CREATION_TIME, MODIFICATION_TIME, PUBLICATION_TIME, RELEASE_TIME, RECEPTION_TIME, EXPIRATION_TIME o NONE
beginPoint y endPoint	Se usa cuando se etiqueta una duración, que está determinada por otras expresiones temporales
quant y frec	Se usa para cuantificar una expresión temporal de tipo SET

Cuadro 2.6: Atributos de TIMEX3

SIGNAL

Sirve para anotar elementos del discurso que explicitan la relación existente entre dos entidades (dos eventos, dos expresiones temporales o una expresión temporal y

un evento).

Generalmente se trata de:

- Preposiciones: *on, in, at, from, to, before, after, during, etc.*
- Conjunciones: *before, after, while, when, etc.*
- Caracteres especiales: por ejemplo "-" y "/" para denotar rangos (*May 12-17*)

Su único atributo es un identificador.

Ejemplo:*John left two days before yesterday*

John left two days <SIGNAL sid="s1">**before**</SIGNAL> yesterday

MAKEINSTANCE

Se crea para marcar cada realización de un evento, ya que puede haber casos en los que se producen varias. Por ejemplo, en la oración:

"John taught on Monday and Wednesday"

El evento "*teach*" sucede en dos ocasiones, "*on Monday*" y "*on Wednesday*". Por tanto habría dos realizaciones de un mismo evento.

Además del ID de la instancia y del evento, posee atributos que capturan la categoría gramatical del verbo, así como la categoría sintáctica de la frase etiquetada como evento. Estos atributos son:

- **tense:** PRESENT, PAST, FUTURE, INFINITIVE, PRE_PART, PAST_PART o NONE.

- **aspect:** PROGRESSIVE, PERFECTIVE, PERFECTIVE_PROGRESSIVE o NONE

- **POS:** ADJECTIVE, NOUN, VERB, PREP

- **Polarity:** valor booleano que indica si la instancia del evento es una negación

- **Modality:** presente si hay una palabra modal que modifica la instancia

Ejemplo: *John left two days before yesterday*

John <EVENT eid="e1" class="OCCURRENCE">**left**</EVENT> two days before yesterday

<MAKEINSTANCE eiid="ei1" eventID="e1" tense="PAST"

aspect="PERFECTIVE"/>

LINK

Esta etiqueta es la gran novedad respecto a los esquemas anteriores. Sirve para representar las relaciones entre los eventos, *"sacando"* esta información de dentro del etiquetado de dichos eventos.

Hay tres tipos de etiquetas:

- **TLINK** Representa la relación temporal entre eventos, momentos o eventos y momentos. Se contemplan los tipos: BEFORE, AFTER, INCLUDES, IS_INCLUDED, DURING, DURING_INV, SIMULTANEOUS, IAFTER, IBEFORE, IDENTITY, BEGINS, ENDS, BEGUN_BY y ENDED_BY.

  Ejemplo: *John left two days before yesterday*

  John <EVENT eid="ei1" >**left**</EVENT> two days before yesterday

  <TLINK eventInstanceID="ei1" relatedToTime="t3" signalID="s1"

  relType="IS_INCLUDED"/>

- **SLINK** Representa la relación temporal entre un evento y otro subordinado sintácticamente a él. Se contemplan los tipos: MODAL, FACTIVE, COUNTER_FACTIVE, EVIDENTIAL, NEGATIVE EVIDENTIAL y CONDITIONAL.

  Ejemplo: *Mary said John left two days before yesterday*

  Mary <EVENT eid="e1">**said**</EVENT> John <EVENT eid="e2">**left**

  </EVENT> two days before yesterday

  <MAKEINSTANCE eiid="ei1" eventID="e1"/> <MAKEINSTANCE eiid="ei2" eventID="e2"/>

  <SLINK eventInstanceID="ei1" subordinatedEventInstance="ei2"

  relType="EVIDENTIAL"/>

- ALINK Representa la relación entre un evento de tipo aspectual y su argumento.

Ejemplo: *John started to read*

John <EVENT eid="e1">**started**</EVENT> to <EVENT eid="e2">**read**

</EVENT>

<MAKEINSTANCE eiid="ei1" eventID="e1" /> <MAKEINSTANCE eiid="ei2" eventID="e2" />

<ALINK eventInstanceID="ei1" relatedToEventInstance="ei2"

relType="INITIATES" />

TimeBank

El corpus **TimeBank** [35] fue **manualmente anotado** con el lenguaje **TimeML** en el marco del Congreso **TERQAS**, en el que éste había nacido. El objetivo era que sirviera de recurso tanto a los investigadores de corpus lingüísticos interesados en Lenguaje y Tiempo como a los desarrolladores de aplicaciones de Extracción de Información, Búsqueda automática de respuestas, etc. para los que el conocimiento de la posición y el orden de eventos en el tiempo es de vital importancia.

Se componía de 183 artículos de noticias (unos 61000 *tokens*) procedentes de diversas fuentes: CNN, New York Times o Wall Street Journal.

Como veremos a continuación, TimeBank ha servido de Gold Standard en varias tareas competitivas.

2.2.2. La tarea de Anotación Temporal a través de los Congresos y Foros de Evaluación

En los últimos veinte años, el interés de la comunidad investigadora por mejorar el manejo de la información temporal por parte de todo tipo de sistemas, no ha hecho más que crecer. Como consecuencia de ello, se han celebrado numerosos e importantes Congresos y Foros de Evaluación sobre extracción y normalización de expresiones temporales.

En la sexta edición de la **MUC** (Message Understanding Conference) [32], en 1995, se introdujo el reconocimiento de expresiones temporales como una sub-tarea del Reconocimiento de Entidades Nombradas. Se requería identificar dos tipos de expresiones, fechas y horas, sin que fuera necesario proceder a su normalización a

ningún estándar de tiempo. En la edición siguiente, **MUC-7** [14], volvió a incluirse la tarea, ampliando el alcance del marcado de las expresiones a los modificadores que podían acompañarlas y añadiendo la identificación de expresiones relativas e implícitas (en concreto, nombres de periodos festivos).

Como consecuencia del interés de la comunidad investigadora en Sistemas de Búsqueda Automática de Respuesta (QA) en la explotación de la información temporal, se celebró, en 2002 el congreso **TERQAS** [33] (Time and Event Recognition for Question Answering Systems). Organizado por **James Pustejovsky**, tenía como propósito dotar a los sistemas QA, de la capacidad de contestar, adecuadamente, preguntas de contenido temporal sobre los eventos y entidades descritos en artículos de noticias. Se fijaron dos objetivos: tratar de distinguir, formalmente, eventos y su realidad temporal en artículos de noticias y evaluar y desarrollar un algoritmo para la identificación y extracción de eventos y expresiones temporales de un texto. Para ello se definió un lenguaje de especificación para eventos y expresiones temporales, que fue la primera version de **TimeML** y se anotó con él un corpus para que sirviera de Gold Standard temporal (**TimeBank**).

Entre 2004 y 2007, se incluyó en la **ACE** (Automatic Content Extraction), una tarea combinada de extracción y normalización de expresiones temporales. Los corpus que se habían de procesar estaban anotados con respecto a las guías de anotación TIDES.

También en 2007 se organizó por primera vez, dentro del Workshop **SemEval** [49], el desafío **TempEval**. Se proporcionaba a los participantes textos del corpus **TimeBank**, anotados con información relativa a expresiones temporales y eventos[34] y se proponía la tarea de determinación de las relaciones temporales entre eventos y fecha de creación, entre eventos y expresiones temporales y entre eventos entre si. En la segunda edición, celebrada en 2010 se añadieron dos tareas de reconocimiento y extracción, respectivamente de expresiones temporales y eventos. Así, se podía afrontar la tarea completa de identificación de todas las expresiones y relaciones con algún grado de temporalidad. El descubrimiento de relaciones entre dos eventos se dividía en dos tareas: la identificación de relaciones entre dos eventos principales (en ocasiones consecutivos) y la identificación de relaciones entre dos eventos sintácticamente dependientes. Mientras en la primera edición las tareas fueron ofrecidas solo

para el idioma inglés, en la segunda se amplió la oferta a seis lenguas, aunque solo se presentaron sistemas para el tratamiento de textos en español e inglés.

Ya en 2013 se celebró la tercera edición [48], cuya particularidad fue que los corpus proporcionados para las tareas de clasificación de las relaciones, estaban formados por texto libre, sin anotaciones; así, los participantes tenían que extraer sus propios eventos y expresiones temporales primero, determinar cuáles estaban relacionados y cuál era el tipo de relación. Los idiomas a tratar eran español e inglés.

El reto de SemEval para 2015 mira de nuevo hacia los sistemas de extracción automática de respuesta. Se propone la tarea "**QA TempEval**"[3] en la que se pedirá a los participantes que anoten, con TimeML, un conjunto de documentos de texto plano y se tratará de evaluar en qué grado contribuyen dichas anotaciones a que los sistemas QA comprendan la información temporal del texto y obtengan respuestas correctas a las preguntas planteadas. Así, en vez de evaluar la calidad de la anotación, comparándola con una anotación manual, se contará el número de respuestas correctas obtenidas por el sistema QA, apoyándose en la anotación.

2.2.3. Anotación automática

Se ha hecho mucha anotación manual, sobre todo para la construcción de corpus como TimeBank, pero puesto que el objetivo de la anotación temporal es enriquecer el procesamiento de grandes cantidades de información es imprescindible que esta anotación se haga de forma automática.

De los distintos Foros de Evaluación, han resultado varias herramientas de rendimiento notable y que, en general, están a disposición de la comunidad investigadora para ser usadas libremente, con fines académicos, como TARSQI, HeidelTime o Terseo.

Terseo [40], es un sistema capaz de reconocer y normalizar información temporal, usando dicha información para ordenar los eventos presentes en un texto. Tiene una base de datos de expresiones y reglas de resolución, que aunque en principio solo contenía normas para el español, se ha ido extendiendo a otros idiomas (inglés e italiano) mediante la adquisición automática de nuevas reglas. Lamentablemente no es accesible en este momento.

[3] http://alt.qcri.org/semeval2015/task5/

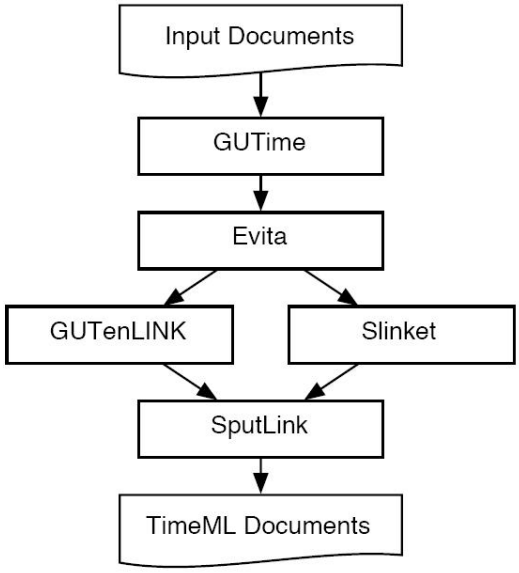


Figura 2.2: Arquitectura de Tarsqi-TTK [50]

En el desarrollo de nuestra propuesta nos serviremos de **HeidelTime** para la identificación de expresiones temporales en un conjunto de *tweets* en inglés y español; también utilizaremos **TARSQI** con el subconjunto de los *tweets* en inglés, por su capacidad para detectar eventos. A continuación haremos un repaso de las características más reseñables de estos sistemas.

TARSQI-TTK

TARSQI [50] es un sistema modular para la anotación temporal de textos de noticias. Está formado por una serie de subprogramas que van añadiendo distintas anotaciones al documento (ver Figura 2.2):

- *GUTime* extiende las capacidades del primitivo anotador **TempEx**, desarrollado por Mani y Wilson [29] para TIMEX2 y que reconocía y normalizaba expresiones temporales, tanto absolutas como relativas. En concreto GUTime adapta TempEx al nuevo estándar TIMEX3 y anota duraciones y varios modificadores temporales.

- *Evita* es un sistema de reconocimiento de eventos. Localiza y anota todas las expresiones referidas a eventos que pueden ser temporalmente ordenados. Para ello, identifica aspectos gramaticales como el tiempo verbal, la polaridad, la modalidad o la clase de los eventos.
- *GUTenLINK* anota las relaciones entre eventos con la etiqueta TLINK. Para ello usa reglas manuales y léxicas.
- *Slinket* identifica construcciones subordinadas que introducen información de modalidad en el texto, tal como vimos en la definición de la etiqueta SLINK, al describir TimeML.
- *SputLink* analiza las relaciones temporales detectadas e infiere nuevas relaciones, aplicando reglas de transitividad, y a la manera en que se hacía en la representación mediante grafos de la información temporal presente en un documento (ver Apartado 2.1.1). Este mecanismo, casi cuadruplica las relaciones existentes en un documento anotado con TimeBank.

Se puede acceder libremente a TARSQI, solicitándolo a través de la web de TimeML[4]. Cuenta con una interfaz gráfica (ver Imagen 2.3) y es bastante manejable; su principal inconveniente es que sólo anota textos en inglés.

HeidelTime

HeidelTime [47] es un anotador temporal **multilingüe y multidominio** que extrae y normaliza expresiones temporales, usando el estándar TIMEX3. Se usan diferentes estrategias según el género de los textos tratados: noticias, narrativa (para artículos de Wikipedia, por ejemplo), coloquial (para sms, *tweets*,...) y científico.

Es un sistema basado en reglas, con una arquitectura separada para el código principal y los recursos lingüísticos (patrones y reglas de normalización), por lo que se puede extender fácilmente a otros idiomas.

Puede descargarse mediante suscripción, a través de la web de la Universidad de HeidelBerg [5]. También está disponible una demo online[6] (ver Figura 2.4).

[4] http://timeml.org/site/tarsqi/toolkit/download.html
[5] http://dbs.ifi.uni-heidelberg.de/index.php?id=form-downloads
[6] http://heideltime.ifi.uni-heidelberg.de/heideltime/

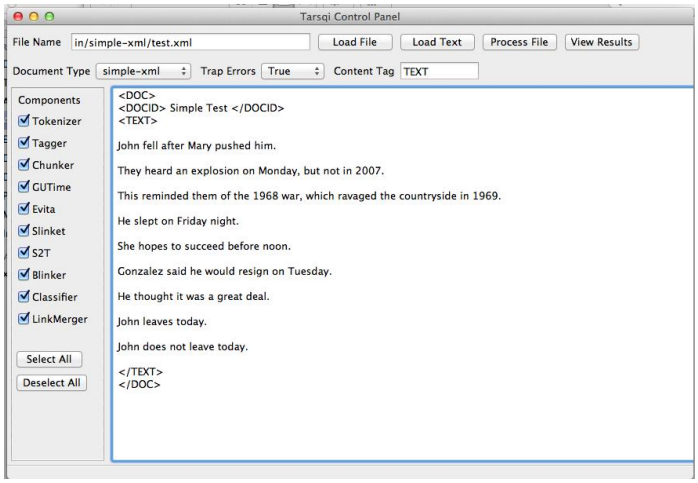


Figura 2.3: Interfaz gráfica de Tarsqi-TTK

Como dijimos, cuenta con recursos para varios idiomas, entre ellos el español. Está escrito en Java y puede integrarse fácilmente en otros programas. Su inconveniente principal es que sólo etiqueta expresiones temporales y no eventos.

este caso, resaltando el aspecto temporal) y aparecen en la lista de resultados de una consulta para facilitar la búsqueda de información **Alonso et al.** [4].

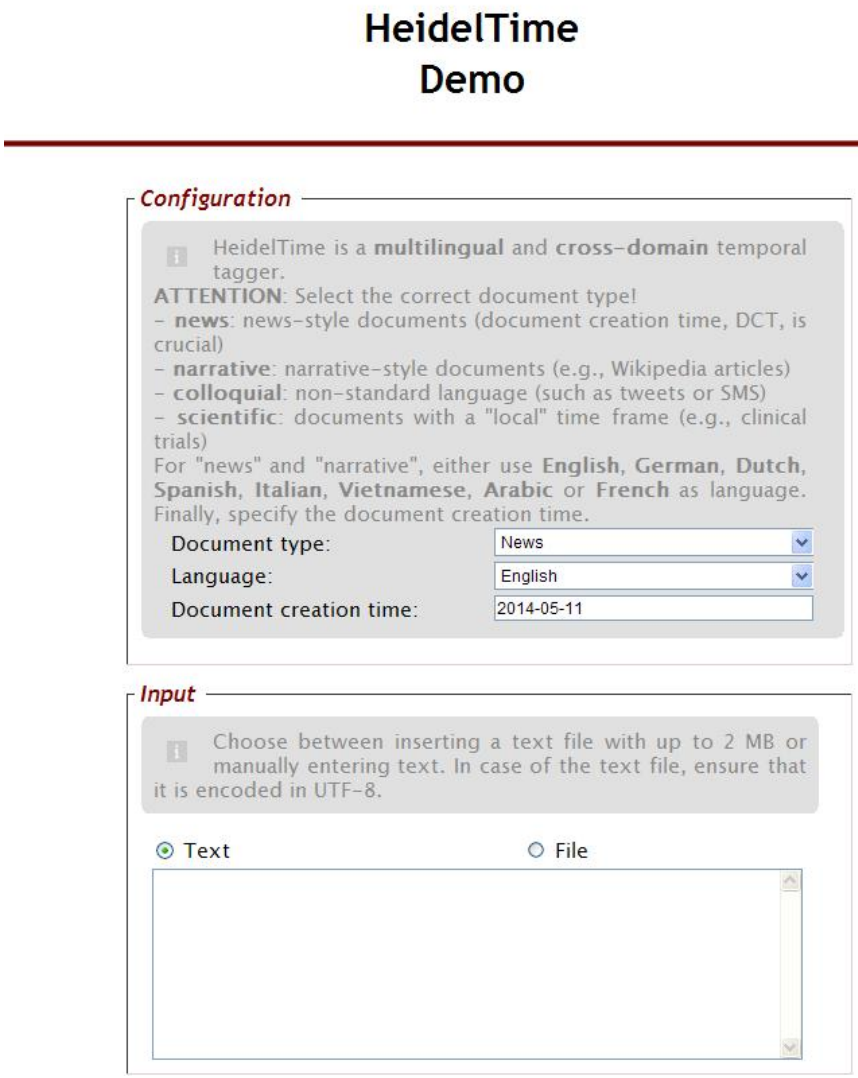


Figura 2.4: Demo online de HeidelTime

Capítulo 3

Análisis Formal de Conceptos. Aplicación a la organización de resultados de búsqueda

El Análisis Formal de Conceptos (en adelante, AFC) es una rama de la Matemática Aplicada, desarrollada por **Rudolf Wille** en 1984, que trata de la formalización de conceptos y de su organización en una estructura de retículo (o jerarquía conceptual). Un concepto formal es una abstracción del pensamiento humano que permite interpretar la información de un modo significativo. La capacidad de los retículos para organizar los conceptos y hacer explícitas las relaciones entre ellos, convierte el AFC en una poderosa herramienta susceptible de ser utilizada en tareas de Recuperación de Información.

En este Capítulo presentaremos los fundamentos de la teoría de AFC y veremos en qué forma se puede aplicar a la organización de resultados de búsqueda.

3.1. Fundamentos matemáticos del AFC

Contextos formales

Un **contexto formal** es una terna $K := (G, M, I)$ donde: $G = \{g\}$ es el conjunto de **objetos**, $M = \{m\}$ es el conjunto de **atributos** que se pueden aplicar sobre los

objetos e $I \subseteq G \times M$ es una relación binaria que relaciona los objetos de G con los atributos de M que poseen o se les aplican.

Ejemplo:

Sea $G = \{1, 2, ..., 10\}$, el conjunto de los diez primeros números naturales y sea $M = \{par, impar, primo, compuesto, cuadrado\}$, una serie de atributos que se aplican sobre los conceptos de G en la forma que indica el Cuadro 3.1 (cuando un objeto g posee un atributo m (gIm), en la matriz de incidencia se indica una "x"). $K = (G, M, I)$ es un contexto formal.

	par	impar	primo	compuesto	cuadrado
1		x			x
2	x		x		
3		x	x		
4	x			x	x
5		x	x		
6	x			x	
7		x	x		
8	x			x	
9		x		x	x
10	x			x	

Cuadro 3.1: Matriz I - Relación de incidencia entre $G = \{1, 2, ..., 10\}$ y $M = \{par, impar, primo, compuesto, cuadrado\}$

Intuitivamente, podemos encontrar que hay objetos que comparten ciertos atributos, y además, son los únicos objetos que los poseen. Por ejemplo los objetos del subconjunto $\{3, 5, 7\}$ tienen en común sus dos atributos, $\{impar, primo\}$ y no hay ningún otro objeto que posea estos dos atributos. Esto nos lleva a la definición clave del AFC, la de concepto formal.

Conceptos formales

Sea $K := (G, M, I)$ un contexto formal, sea $A \subseteq G$ y $B \subseteq M$, (A, B) es un **concepto formal**, y se denota $C(A, B)$ si y sólo si: $A' = B \wedge B' = A$ siendo:

$A' = \{m \in M / gIm \ \ \forall g \in A\}$ *(atributos de M que aplican sobre todos los objetos de A)*

$B' = \{g \in G / gIm \ \forall m \in B\}$ *(objetos de G sobre los que aplican todos los atributos de B)*

A es la **extensión** del concepto y B su **intensión**.

Dado un contexto formal, es posible encontrar todos sus conceptos formales, según se establece en [22]:

"Cada concepto de un contexto (G, M, I) es de la forma (X'', X') para algún $X \subseteq G$ y de la forma (Y', Y'') para algún $Y \subseteq M$. Recíprocamente, todos los pares construidos de esta forma, son conceptos."

Los conceptos que están generados por un único objeto ($\gamma(g) = (g'', g)$) o un único atributo ($\mu(m) = (m', m'')$) se denominan, respectivamente, **concepto-objeto** y **concepto-atributo**.

Volviendo al ejemplo, consideremos por ejemplo el objeto $g = \{6\}$. Sus atributos son: *par* y *compuesto*, luego $g' = \{par, compuesto\}$. Para encontrar g'' hemos de buscar el resto de números entre el 1 y el 10 que poseen estos dos atributos, que son: 4, 8 y 10. El concepto objeto es $\gamma(6) = (\{4, 6, 8, 10\}, \{par, compuesto\})$

Retículo de conceptos. Teorema Fundamental del AFC

En el conjunto de conceptos del concepto K, $\beta(K)$, se define una relación de orden ($\leq$), de la siguiente manera:

Dados $C_1(A_1, B_1)$ y $C_2(A_2, B_2)$, se dice que $C_1 \leq C_2$, si $A_1 \subseteq A_2$ (equivalentemente, $B_2 \subseteq B_1$).

Si $C_1 \leq C_2$, C_1 es un **subconcepto** de C_2 (o C_2 es un **superconcepto** de C_1).

$(\beta(K), \leq)$ es un conjunto de ordenado y puesto que se puede comprobar que para todo subconjunto de $\beta(K)$ existen el supremo y el ínfimo, $\beta(K)$ es un retículo. De hecho es un retículo completo, según establece el **Teorema fundamental de los retículos de conceptos**[22]:

" Dado K, contexto formal, $\beta(K)$, el conjunto de todos los conceptos formales es un retículo completo en el que el supremo y el ínfimo son:

$$\bigwedge_{t \in T} (A_t, B_t) = \left(\bigcap_{t \in T} A_t, \left(\bigcup_{t \in T} B_t \right)'' \right)$$

$$\bigvee_{t\in T}(A_t, B_t) = \left(\bigcup_{t\in T} A_t, \left(\bigcap_{t\in T} B_t\right)''\right)$$

Representación del retículo de conceptos

Los retículos de conceptos pueden representarse gráficamente mediante un diagrama de líneas o diagrama de *Hasse*. En dichos diagramas, cada nodo representa un concepto, que se etiqueta con su extensión e intensión. Puesto que un objeto puede pertenecer a la extensión de más de un concepto, se toma el convenio de incluirlo sólo en la etiqueta de su concepto-objeto; de igual manera se etiquetará con un atributo sólo su concepto-atributo.

Obviamente, todos los conceptos que sean superconceptos de uno dado contienen a la extensión de éste en la suya, y todos sus subconceptos contienen en su intensión a la intensión de éste.

En la Figura 3.1 vemos la representación de Hasse del retículo de conceptos del ejemplo.

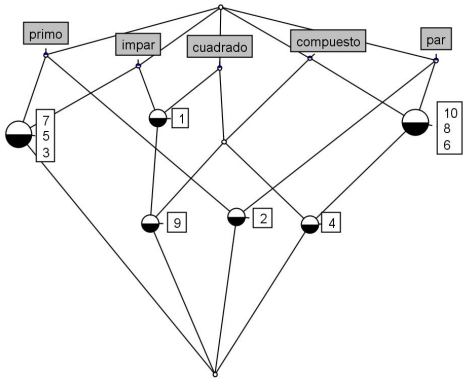


Figura 3.1: Retículo de conceptos (Concept Explorer [56])

3.2. AFC y Recuperación de Información

La capacidad del AFC para organizar la información ha motivado su aplicación a distintas tareas de Recuperación: extensión de las consultas, organización, *clustering*

e incluso alguna forma de detección de temas (*Topic Detection*).

Dada una colección de documentos, G y una serie de términos de búsqueda, M, **Godin** et al. [24] definen el contexto formal $K(G, M, I)$, donde I es la relación dada por el índice de los términos de G en M, Los nodos (conceptos formales) del retículo formado pueden interpretarse como una **consulta** q, cuyos términos son la intensión del concepto y los documentos recuperados son la extensión. Un superconcepto de un concepto dado se traduciría en una **generalización** de la consulta, esto es, al reducir el número de términos (atributos), la cantidad de documentos recuperados (objetos) sería mayor. Inversamente, los subconceptos serían **especializaciones**, añadirían atributos y reducirían la cantidad de documentos recuperados. Esto permitiría asistir al usuario en el proceso de búsqueda, dándole la posibilidad de refinar la consulta o generalizarla mediante la adición o eliminación de atributos.

Ejemplo: Sea un conjunto G de diez documentos relacionados con los *Beatles*, {d1, d2, ..., d10}, y sean los términos de búsqueda M={Beatles, discos, Sgt.Pepper}, donde la aparición de los términos en los documentos viene dada:

d1: Beatles discos

d2: Beatles discos

d3: Beatles discos

d4: Beatles discos Sgt.Pepper

d5: Beatles discos Sgt.Pepper

d6: Beatles discos Sgt.Pepper

d7: Beatles Sgt.Pepper

d8: Beatles Sgt.Pepper

d9: Beatles Sgt.Pepper

d10: Beatles Sgt.Pepper

El retículo de conceptos puede verse en la Figura 3.2.

Los documentos recuperados para la consulta {*Beatles discos Sgt.Pepper* }, por ejemplo, serían d4, d5 y d6; ahora bien, si eliminásramos uno de los parámetros de la consulta, por ejemplo *discos*, obtendríamos también d7,d8 y d9.

Como es sabido, muchas veces hay documentos relevantes para una consulta que no coinciden exactamente con sus términos. Los retículos de conceptos también pueden utilizarse para ordenar los resultados de búsqueda, calculando una suerte de

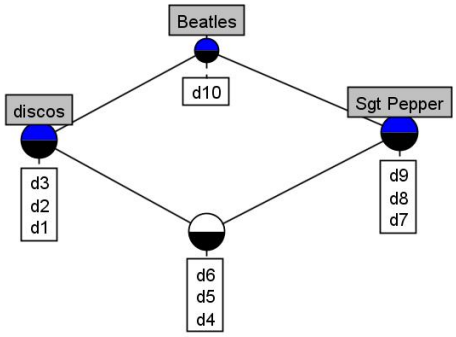


Figura 3.2: Retículo de conceptos para consultas sobre los Beatles

distancia conceptual entre un documento y la consulta, basándose en la estructura del reticulo [12].

3.3. *Clustering* de documentos mediante AFC

Como presenta Cigarrán en su tesis [15], AFC supone una potente alternativa al clustering jerárquico, representando la información de forma más completa y accesible. Esto se debe a que AFC permite respetar dos restricciones: la de un *universo abierto* y la de *herencia múltiple*.

La restricción a un *universo abierto* [15] supone que dada una agrupación de documentos, cada *cluster* contendrá únicamente los documentos que no puedan ser especializados en un cluster más pequeño; así, cada documento aparecerá solo en el cluster que mejor describa sus características principales. En directorios web que utilizan este paradigma, una página web nunca aparece referenciada en distintos niveles jerárquicos, dentro del sistema de categorías.

En cuanto a la *herencia múltiple*, permite que un documento aparezca representado en distintas partes del clustering, cuando presenta características de varios clusters, en contraposición al clustering jerárquico, que trataría de asignar el documento al cluster que mejor lo representara. De esta forma, el usuario podría acceder a un documento por varios caminos, sin perder información.

La adaptación de AFC a estas restricciones, requiere de la definición del concepto

de *"nodo de información"*. Formalmente, dado $C(A, B)$ un concepto formal, su *nodo de información* será un par (AI, BI) donde $AI \subseteq BI$ es el conjunto de elementos de A para los que C es su concepto-objeto y $BI = B$. De la definición se desprende que el nodo de información de un concepto formal contendrá el conjunto de documentos que esté completamente descrito por la intensión del concepto. En un diagrama de *Hasse*, los nodos etiquetados con objetos coinciden, por tanto, con nodos de información: AI será el conjunto de objetos de la etiqueta y BI la intensión del concepto (ver Figura 3.3).

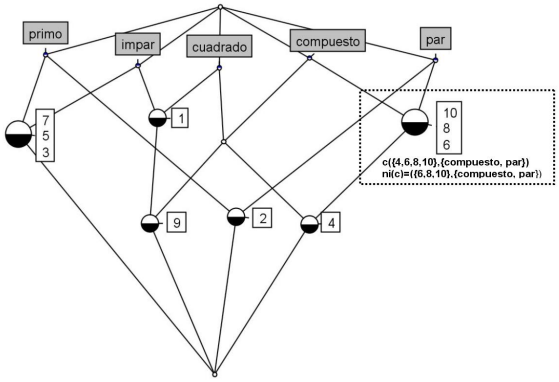


Figura 3.3: Ejemplo de *nodo de información*

Otra de las aportaciones de dicho trabajo es la presentación de una metodología orientada a la construcción del contexto, en cuanto a la obtención de los descriptores o atributos que van a describir los documentos, con el fin de obtener estructuras de clustering manejables y lo menos complejas posible. La extracción de los descriptores se basa en la obtención de unigramas, sintagmas terminológicos y n-gramas; una vez extraídos, se seleccionan de entre ellos aquéllos que describen a los documentos de la forma más completa y significativa, teniendo en cuenta el número de *clusters* generados y el grado de separación entre la información relevante y la no relevante.

3.4. Entornos y herramientas

Existen varias aplicaciones de AFC, muchas de ellas disponibles para uso académico. Se trata de implementaciones de diferentes algoritmos de creación del retículo de

conceptos y generación de las reglas de asociación, representación del diagrama de líneas, etc. Nosotros hemos utilizado **Concept Explorer**[1] [56], que aún no generando las relaciones de forma muy eficiente, tiene buenas propiedades gráficas.

Concept Explorer permite, a través de su interfaz gráfica (ver 3.4) generar el retículo de conceptos, y modificarlo para aumentar su representatividad, desplazando los nodos, mostrando o no mostrando las etiquetas de atributos y objetos, etc. Calcula el número de conceptos y las reglas de asociación y permite reducir el contexto, eliminando atributos.

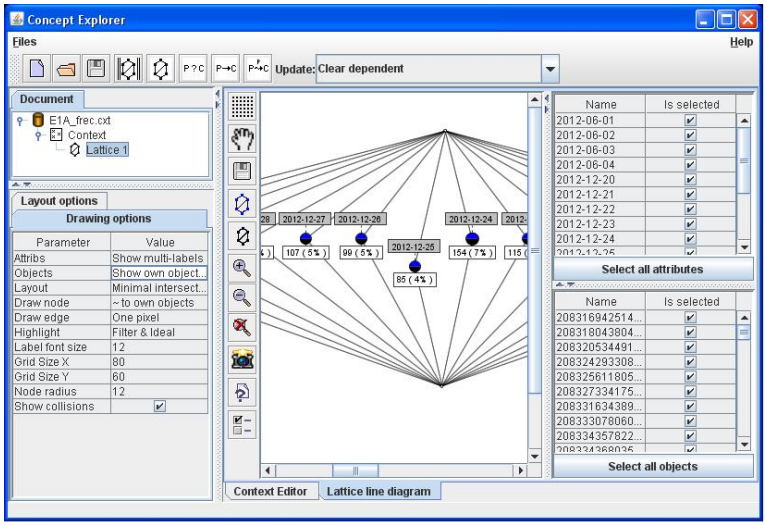


Figura 3.4: Interfaz gráfica de Concept Explorer

También hemos utilizado la herramienta desarrollada por Cigarrán et al. [16], que calcula los conceptos y su estabilidad.

[1] http://conexp.sourceforge.net

Capítulo 4

Enfoques basados en el contenido temporal para la organización de información

En este capítulo se abordan los trabajos preliminares en anotación y explotación de la información temporal asociada a un documento. Con especial nivel de detalle, se describen los trabajos relacionados directamente con nuestra propuesta.

4.1. La explotación de la Información Temporal

Tradicionalmente, los sistemas de Recuperación de Información, han hecho un uso parcial de la información temporal asociada a los documentos. En general, solo la fecha de creación o modificación era tomada en consideración y cuando se atendía a la expresiones temporales incrustadas en el contenido, no se les daba una dimensión temporal, sino que se trataban como simples cadenas de texto.

En los últimos años, sin embargo, con los avances en anotación automática, ha habido una intensa investigación y se han desarrollado un buen número de sistemas que organizan y presentan la información teniendo en cuenta su contenido temporal. En las aplicaciones comerciales, como buscadores, etc., se empieza a hacer habitual la presencia de líneas de tiempo que ayudan al usuario a contextualizar temporalmente los resultados de búsqueda, aún así, como veremos, hay un largo camino por recorrer.

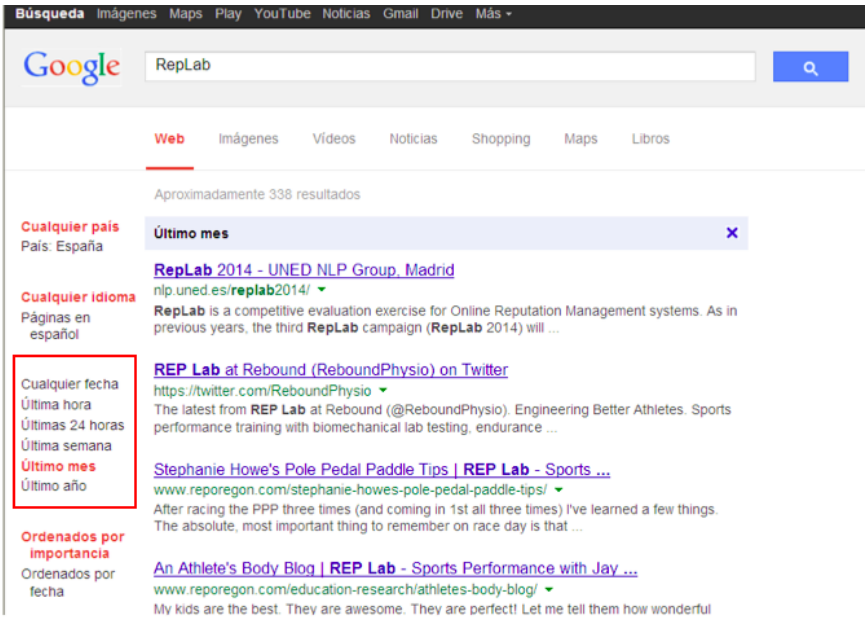


Figura 4.1: Google *timeline*

Recuperación Temporal de Información

El escenario planteado es el de un usuario que tiene una necesidad de información, concreta o general, y busca satisfacerla mediante un sistema que le facilita la tarea, filtrando y organizando la información relevante.

La construcción de líneas de tiempo entorno a las que se organizan los resultados de búsqueda es ya una práctica muy extendida (ver Figura 4.1). Bien es cierto que en general, se utiliza como única referencia la fecha de creación o modificación del documento, y según el dominio, puede no representar adecuadamente el contenido. Por ejemplo, no es lo mismo ordenar por fecha de creación una serie de noticias que artículos de la Wikipedia [1].

La experiencia del usuario en la exploración de resultados de búsqueda mejora mucho cuando se le presentan éstos con un cierto nivel de agrupamiento, en lo que se llaman *hit-lists*. **Alonso** et al. [5] plantean un agrupamiento basado en un *perfil temporal de los documentos*, que permite su ubicación en líneas de tiempo de dis-

[1] Wikipedia: http://wikipedia.org

tinta granularidad. Este perfil temporal está formado por las expresiones temporales (explícitas, implícitas y relativas) que contiene el documento y en base a él, se calcula la medida de similitud que genera los agrupamientos.

También se ha probado la efectividad del uso de *snippets temporales*, pequeños textos que describen el contenido de un documento (en este caso, resaltando el aspecto temporal) y aparecen en la lista de resultados de una consulta para facilitar la búsqueda de información **Alonso et al.** [4].

En el campo de la Búsqueda Web, trabajos como el de **Vicente-Díez** y **Martínez** [53] ponen de manifiesto cómo mejora el rendimiento de los sistemas al considerar la semántica temporal tanto de la consulta como de la colección de documentos. Para ello, utilizan un anotador basado en la etiqueta *TIMEX2* de **TIDES** (ver Apartado 2.2.1). En esta línea, y para consultas de carácter temporal implícito, destaca el trabajo de **Campos et al.** [11].

El tiempo es también el factor clave en la serie de recientes trabajos que, directamente, buscan estudiar cómo ha evolucionado temporalmente un determinado concepto [31], o la relación entre ciertas entidades o eventos [23, 37]. Mientras en [37] el enfoque es *naive* y, una vez identificada una relación, se procede a definir el periodo temporal en que ésta surge, atendiendo solo a la fecha de publicación, en [23] utilizan **TARSQI** (ver Apartado 2.2.3) para extraer expresiones temporales y relacionarlas con los eventos.

Sumarización y *Topic Detection*

El escenario planteado es el de una persona que tiene que *monitorizar* un flujo continuo de noticias, entradas de un *blog*, *tweets*, etc.

Los sistemas que asisten al usuario en este tipo de tareas van encaminados a ahorrarle el procesado de información irrelevante por no ser novedosa, así como a proporcionarle resúmenes periódicos para que no se pierda nada en la corriente de información. Por ello, aúnan técnicas de sumarización y *topic detection* o detección de temas. En realidad, comúnmente se detectarán "*eventos*" y no "*temas*"; ya que por ejemplo un tema sería "*huracán*" mientras que el evento sería "*huracán Katrina*" o "*huracán Irene*".

Allan et al. [1] extraen resúmenes temporales en el dominio de las noticias,

seleccionando una frase de cada evento dentro de un tema. Detectan los cambios que se producen en los eventos de cada tema evaluando su "grado de novedad", para ponderar la importancia de la frase que lo contiene, cara a incluirla en el resumen. Los resúmenes son elaborados a intervalos regulares de tiempo (cada hora, al comienzo de la mañana,...) de forma que reflejen solo el contenido nuevo, sin que tenga sentido hacer un resumen global cuando el tema ha expirado o en cada intervalo, ya que el usuario conoce la información anterior. Dicho trabajo no utiliza la información temporal contenida en las noticias, al contrario que **Makkonen et al.** [28], que explota distintas clases semánticas (lugar, nombres propios, expresiones temporales y términos generales) para evaluar la novedad o no de un evento.

Recientemente, **Arkaitz et al.** [57] se acercan a la **sumarización en tiempo real** en **Twitter**, para eventos programados, esto es, cuya celebración o acontecimiento está fijado y es conocido con anterioridad. En concreto, desarrollan un sistema que monitoriza, analizando el texto, los *tweets* relativos a partidos de fútbol de la Copa América. En dicho contexto hay cierta información previa sobre el tipo de eventos esperados (goles, expulsiones, tarjetas,...) lo que facilita la tarea de detección de sub-eventos. Una vez detectado un sub-evento nuevo, el siguiente paso es seleccionar el tweet que mejor refleje lo que ha pasado.

Análisis temporal de la Web

El escenario en la Web tiene muchas facetas. Obviamente estamos ante una cantidad de información exorbitante y dinámica, cuyo tremendo crecimiento exige formas más y más refinadas de organización y presentación. En el camino hacia la web semántica, el razonamiento temporal de los sistemas es prácticamente un requisito.

Pero además, los motores de búsqueda deben recorrer continuamente la red para indexar el contenido nuevo, así como las modificaciones que se hayan producido. Un conocimiento adecuado de la temporalidad de este contenido es imprescindible para aumentar la eficiencia del *crawling*.

Por otro lado, como consecuencia de la interacción entre los miembros de las redes sociales, se genera un inmenso volumen de datos en *logs*, que con un adecuado procesamiento puede ayudar a entender la dinámica y evolución de las mismas.

Desde 2011 se viene celebrando el Congreso **TWAW** (*Temporal Web Analytics*

Workshop) [9]. Organizado por **Omar Alonso**, su objetivo es *ser un punto de encuentro para investigadores de todos los campos en los que la dimensión temporal de la información abre un nuevo abanico de posibilidades, así como introducir la información temporal en el análisis Web*. En sus distintas ediciones, se han presentado trabajos sobre análisis temporal de redes sociales, archivo Web, datación del contenido e identificación y aprovechamiento de expresiones temporales. En la edición de 2014, que acaba de celebrarse, se presentó un intento de construcción de una ontología temporal, **TempoWordNet** [17], que esperemos sirva para enriquecer la anotación automática y aumentar la independencia del dominio.

4.2. Clustering temporal

Nuestra propuesta se inspira en el trabajo de **Alonso et al.** [5], *"Clustering and exploring search results using timeline constructions"*. Alonso et al. extraen las expresiones temporales explícitas, implícitas y relativas de un documento y las normalizan para construir un *perfil temporal* del mismo. Este perfil temporal se adapta a la granularidad que mejor define a la colección (según el Calendario Gregoriano, *ver Apartado 2.1.2*) y se procede a su representación 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 correspondan a más de un *cluster*, cuando tienen varias expresiones temporales; en este caso, tratan de determinar el "*cluster principal*", atendiendo a la expresión temporal predominante (que aparezca más en el documento). El número de documentos en los *clusters* puede ser muy variable, *clusters* con muchos documentos corresponderán a momentos álgidos para el tema en cuestión.

Nuestra idea de representación es similar en cuanto a la información temporal utilizada, pero AFC presenta dos ventajas fundamentales: no requiere seleccionar un descriptor temporal principal, como consecuencia de la restricción de herencia múltiple y el uso simultáneo de varias granularidades se deduce naturalmente de la estructura del retículo. Además, introduciremos la noción de *Calendario Imaginario-Colectivo*; en dicho calendario representaremos todas esas expresiones temporales que no están completamente determinadas, sino que pueden pertenecer a un año

cualquiera, como el mes de *junio* o el *1 de Mayo*. Lo explicaremos en detalle en el siguiente apartado.

Por último y para el caso concreto de Twitter, Alonso et al. no utilizan más información temporal que la fecha de creación de los *tweets*. En base a ésta realizan agrupaciones que ponen de manifiesto cómo ha evolucionado el interés de un determinado tema (ver Figura 4.2). Obviamente, dicha agrupación sólo tiene sentido para consultas muy especificas y no aporta información relevante si el objeto de la búsqueda es una entidad en genérico (como en el caso de la colección RepLab, que luego describiremos). Nuestro modelo busca aplicarse en toda su extension también a este dominio, de hecho, la definicion de *Calendario Imaginario-Colectivo* está inspirada en la habitual conmemoración de aniversarios y fechas señaladas, por los usuarios de las redes sociales.

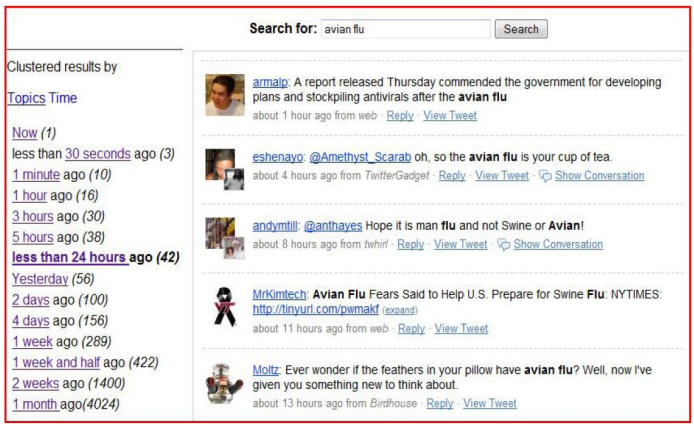


Figura 4.2: Línea de tiempo para *tweets* sobre "avian flu" [5]

4.3. Modelando el contenido de los *tweets*

El otro trabajo de referencia es el de **Castellanos et al.** [13], *"Modelling Techniques for Twitter Contents: A step beyond classification based approaches"*. Este artículo resume la participación del grupo de Procesamiento de Lenguaje Natural de la UNED en RepLab 2013[2], en concreto en las tareas de Polaridad, Filtrado y

[2]http://www.limosine-project.eu/events/replab2013

Detección de temas o *Topic Detection*. Castellanos et al. se enfrentan a esta última tarea mediante un enfoque basado en AFC, que modela los *tweets* tomando como descriptores sus términos. Pretenden así sacar partido del conjunto de entrenamiento, pues FCA permite de forma natural la adaptación a nuevos temas; cosa que no ocurre con las aproximaciones tradicionales basadas en clasificación.

La selección de descriptores basada en el contenido de los *tweets* 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. Para encontrar los *clusters* más susceptibles de representar un tema, se aplica el concepto de **estabilidad**. Dicho concepto, de la disciplina del FCA, indica cuánto depende la intensión de un concepto de un objeto particular de su extensión; un valor alto de estabilidad, medida como la probabilidad de que un concepto conserve su intensión al eliminar un número arbitrario de objetos de su extensión, indica que el concepto contiene un conjunto de *tweets* cohesivo, que puede representar un *cluster*. En nuestra propuesta trataremos de solventar estos problemas definiendo un conjunto de descriptores más reducido y aumentando la independencia del dominio.

Parte II

Propuesta y experimentación

Capítulo 5

Propuesta de trabajo

5.1. Introducción

Como hemos motivado en la primera parte del trabajo, en una colección de documentos hay un montón de información temporal. La extracción de dicha información y su contextualización, abre un amplio abanico de posibilidades de integración de la dimensión temporal en tareas de organización y representación de los documentos.

AFC se ha postulado como una potente herramienta para la organización de resultados de búsqueda. Como vimos en el apartado anterior, se basa en la selección de un conjunto de descriptores que caracterizan una colección de documentos para la agrupación de éstos en una serie de *clusters*.

Nuestra propuesta trata de explotar la información temporal de los documentos para proceder a su agrupación mediante AFC. Esto es, los descriptores que utilizaremos tendrán una dimensión temporal. Se trata de explorar alternativas a las líneas de tiempo, que ahonden en la búsqueda de similitudes entre los documentos, sea cuál sea su dominio de procedencia (noticias, artículos de la Wikipedia, *tweets*,...).

5.2. Propuesta de un modelo para la explotación y organización de la información temporal

Sea $\Delta = \{d_1, ..., d_n\}$ una colección de documentos, su información temporal puede ser de los siguientes tipos:

- **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.

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

Es importante notar que todos los elementos de τ son distintos, esto es, cada evento y cada expresión temporal sólo aparecerán reflejados una vez, aunque haya varios documentos en los que aparezcan o lo hagan varias veces en un mismo documento.

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.

Construcción de τ

Los elementos de τ son nuestro conjunto de *descriptores*, si adoptamos la terminología de Cigarrán [15] y Castellanos [13].

Para su construcción, en concreto, para la normalización de sus expresiones temporales, tenemos que fijar un calendario. Lo tradicional es utilizar el calendario **Gregoriano**. En dicho calendario, como vimos en el apartado 2.1.2, las granularidades serían:

$\varrho = (G_{\text{año}}, G_{mes}, G_{\text{día}}, G_{hora}, G_{minuto}, G_{segundo})$

con las relaciones "$\gg$" (más *gruesa*) y "$\ll$" (más *fina*) :

$G_{\text{año}} \gg G_{mes} \gg G_{\text{día}} \gg G_{hora} \gg G_{minuto} \gg G_{segundo}$.

La elección de una granularidad concreta no es necesaria para AFC, 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 documentos 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 AFC, podemos elegir el conjunto de descriptores sin que se produzca ninguna pérdida de información, ni 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.

Sin embargo, 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 importantes para nosotros pues en ciertos dominios, como las redes sociales, 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.

Dependencia del dominio y reducción del contexto formal

El uso de eventos, que como hemos dicho, son expresiones lematizadas de verbos, etc., presentes en los documentos, nos ancla al dominio y aumenta la complejidad de nuestro contexto formal, equiparando nuestro enfoque al de Castellanos [13].

Para eliminar esta dependencia y obtener un contexto más sencillo tenemos varias opciones. La primera de ellas sería reducir el conjunto de descriptores, eliminando totalmente los eventos. Hay que tener en cuenta que, en general, las expresiones temporales no son muy abundantes, por lo que puede que muchos documentos quedaran descritos solo por su fecha de creación. Otra opción, sería describir cada documento en función del tipo de eventos que contiene (de ocurrencia, de percepción, etc...). De esta forma ganaríamos independencia del dominio, conservando la información en cierta forma y sólo con siete atributos, esto es, con una cantidad más que manejable de descriptores.

En el apartado de Experimentación, realizaremos pruebas con las distintas configuraciones de τ que hemos mencionado.

5.3. Desarrollo del entorno computacional

Para llevar a cabo los experimentos que se describen en el próximo capítulo, nos hemos servido de los anotadores automáticos Heideltime y Tarsqi y de las herramientas de Análisis Formal de Conceptos, Concept Explorer y Cigarrán et al. [16]. Ya hemos hablado de estas herramientas en el Estado del Arte, pero aquí queremos detallar la arquitectura del entorno computacional desarrollado para su integración

y uso. Se tratarán además aspectos básicos de instalación, configuración, formato de los datos de entrada y salida, etc.

El esquema general seguido se representa en la Figura 5.1.

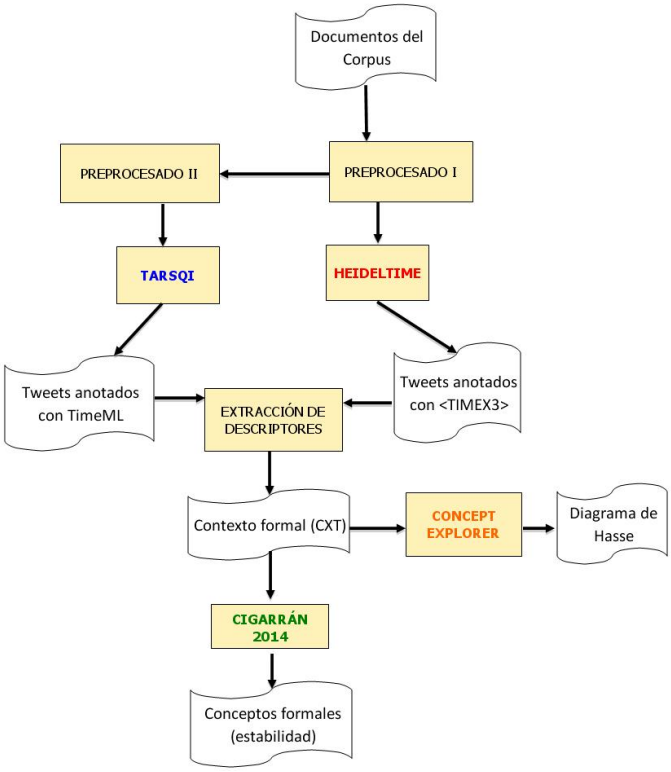


Figura 5.1: Arquitectura del desarrollo computacional de nuestra propuesta

5.3.1. Preprocesado de los datos

Puesto que la anotación, tanto de Heideltime como de Tarsqi, genera un documento conforme a XML, la primera precaución que debemos tomar es eliminar del contenido de los *tweets* los símbolos no permitidos por dicho estándar: <, >, &, ", '. Observamos que con frecuencia se utiliza la combinación "< 3", que se corresponde con un emoticón en forma de corazón. Así pues, al eliminar el símbolo "<", hemos de tenerlo en cuenta, ya que si no eliminamos también el *"3"*, corremos riesgo de que sea anotado.

Programamos un script (**preprocesado.py**) en el que limpiamos los *tweets* de esos caracteres y además eliminamos las *urls*, ya que observamos que es habitual que aparezcan años en las rutas de carpetas, y considerando que la fecha de almacenamiento de un contenido no necesariamente lo contextualiza, queremos evitar su anotación.

Nos referimos a este tipo de *tweets*:

*"The Beatles http://www.globalmusichistory.com/**2012/01**/beatles.html..."*

El anotador Tarsqi procesa archivos en varios formatos, incluido texto plano, pero por las razones que explicaremos a continuación, tenemos que realizar la conversión a formato TimeML de la colección de *tweets*. El script **toTarsqi.java** genera un archivo TimeML para cada *tweet* de la colección **en inglés** (recordemos que no etiqueta textos en español).

5.3.2. Anotación

Tarsqi está desarrollado en Python y no tiene versión para Windows, aunque según los desarrolladores puede ser utilizado en dicha plataforma. Nosotros lo utilizamos en Macintosh con buenos resultados.

Hay que seguir cuidadosamente las instrucciones de instalación, puesto que Tarsqi precisa del procesador **POS TreeTagger**[1] [44]; hemos de descargarlo y colocarlo en la ubicación exacta, tal como explica el manual[2] [51].

El interfaz gráfico sirve para realizar algunas pruebas, pero las ejecuciones masivas han de hacerse a través de línea de comandos.

Tarsqi, como sistema utilizado en Foros de Evaluación, está preparado para recibir archivos pertenecientes a algunos de los corpus más usados en dichos foros, como Timebank. Puede recibir cuatro tipos de archivos, siendo necesario indicar el tipo escogido. Esto presenta ciertos problemas a la hora de **resolver las expresiones relativas**, pues no en todos los formatos se le puede facilitar una fecha de referencia y la forma de hacerlo cambia de unos formatos a otros. Tomando de ejemplo los corpus de prueba, adoptamos el formato de una colección de noticias del NYTimes, esto es importante, porque fue necesario incluso renombrar los identificadores de los *tweets*,

[1] http://www.cis.uni-muenchen.de/ schmid/tools/TreeTagger/
[2] http://www.timeml.org/site/tarsqi/toolkit/manual

para que reconociera la fecha, que además tenía que estar en formato anglosajón (ver Figura 5.2). Una prueba más de la necesidad de utilizar estándares.

Por supuesto tuvimos que dividir la colección de *tweets* y crear un archivo para cada uno, ya que precisábamos ir cambiando la fecha de referencia, pero también mantener el identificador del *tweet* para ligarlo a las anotaciones, y por esta razón no podíamos agrupar *tweets* publicados en la misma fecha.

```
<TimeML xmlns:xsi='http://www.w3.org/2001/XMLSchema-instance'
xsi:noNamespaceSchemaLocation='../../dtd/timeml_1.2.1.xsd'>
<DOCNO>NYTweetID208280017162092544</DOCNO>
<DOCTYPE SOURCE='newswire'> NEWS STORY </DOCTYPE>
<DATE_TIME>05/31/2012</DATE_TIME>
<TEXT>The Beatles were a british boy band who played their own
 instruments and wrote their own songs. i think thats pretty neat</TEXT>
</TimeML>
```

Figura 5.2: Ejemplo de archivo de entrada (TARSQI)

HeidelTime está escrito en Java y es modulable, permitiendo su integración en otros proyectos. De nuevo se necesita instalar y configurar adecuadamente TreeTagger.

La integración en otros proyectos pasa por importar el paquete **HeidelTime Standalone**, tal como se explica en el manual de instalación, e instanciar un objeto *HeidelTimeStandalone*, para lo que hay proporcionarle el idioma y el tipo de documento. Puesto que nuestro corpus era multilingüe, observamos que una vez instanciado el objeto, no es posible cambiar el idioma (aunque existe un método para ello), por lo que tuvimos que instanciar dos, uno para inglés y otro para español. El método *process* de la clase *HeidelTimeStandAlone* tiene como argumentos el texto a etiquetar y la fecha de referencia, por lo que se adaptó perfectamente a nuestros requerimientos.

Integramos HeidelTime en el script Anotador.java, que genera un archivo con todos los *tweets* en formato TimeML, anotados con la etiqueta <TIMEX3>.

5.3.3. Extracción de descriptores y construcción del contexto formal

Para llevar a cabo la batería de experimentos que presentamos en la propuesta, desarrollamos una serie de scripts (E1.py, E2.py,...) que:

- leen los archivos de datos y los archivos de salida de Tarsqi y HeidelTime

- extraen las fechas de creación, las expresiones temporales, los eventos y su tipo, asociadas a cada tweet,

- descartan expresiones poco frecuentes o indeseadas,

- enriquecen las expresiones, añadiendo su equivalencia en otras granularidades de los Calendarios definidos en la propuesta,

- lematizan los eventos usando la librería **nltk.stem.wordnet**[3] [10]

- y finalmente generan la tabla que representa al contexto formal, constituido por los *tweets* (objetos) y sus descriptores temporales (atributos).

La fecha de creación de los *tweets* se suministra en formato UNIX [55], esto es, una fecha dada se representa como la cantidad de milisegundos que han transcurrido entre las 00 horas del 1 de enero de 1970 y dicha fecha. Así pues, ha de efectuarse la conversión al estándar IS0-8601 (ver Apartado 2.1.3), en concreto, al formato *"YYYY-MM-DD"*.

Una vez contruido el contexto formal, recurrimos a los entornos de Análisis Formal de Conceptos, Cigarrán et al. [16] y Concept Explorer, para el cálculo del conjunto de conceptos formales y la representación del retículo.

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

Capítulo 6

Experimentación y evaluación

6.1. Descripción de la colección de prueba

El corpus de experimentación es un subconjunto de la colección de *tweets* de **RepLab2013**. En este apartado se procede a su descripción detallada.

6.1.1. Twitter como corpus

Un *tweet* es un comentario de la red social de noticias **Twitter** [1]. La rapidísima difusión de la información y su impacto social es tal, que empresas, gobiernos y personajes públicos en general, están muy atentos a *lo que se dice de ellos*, esto es , a lo que se denomina su "**reputación online**". Además, Twitter se ha usado como plataforma de algunas formas de democracia participativa, o para organizar eventos masivos de protesta, etc.; en otras ocasiones, son asuntos aparentemente triviales los que alcanzan una espectacular relevancia en un periodo de tiempo muy corto, en lo que se denominan "fenómenos virales". Las tendencias (*trending topics*) de Twitter son uno de los aspectos más estudiados, sobre todo en cuanto a su categorización y predicción.

Si bien desde un punto de vista computacional, es necesario conocer ciertos aspectos de los *tweets* para poder tratarlos adecuadamente, está fuera del alcance de este trabajo profundizar más en los aspectos generales de esta red, por lo que remitimos

[1] http://www.twitter.com

al lector interesado a la bibliografía [8, 26, 27].

Idiosincrasia de los *tweets*

Lo más característico de los *tweets*, además de su asombrosa heterogeneidad, es que su longitud máxima es de 140 caracteres y puede incluir varios de estos elementos:

- *Hashtags*: son etiquetas que se refieren a un tema y se identifican precedidos del símbolo "#". Por ejemplo *#NowPlaying*, es un *hashtag* que se puede utilizar cuando se quiere compartir la canción que se está escuchando en ese preciso momento.
- *Menciones*: se trata de un nombre de usuario precedido por un símbolo "@", sirven para dirigir el contenido del a dicho usuario.
- *Enlaces*: se incluyen para ampliar información o compartir contenido multimedia (imágenes, vídeos). Se suelen mostrar en un formato que reduce su extensión a 20 caracteres.
- *Retweets*: son *tweets* que un usuario comparte en su muro, procedente del de otro usuario, esto es, sin haber sido el autor original. Se identifican con las letras "RT".

La **fecha y hora de publicación** de cada *tweet* (*timestamp*) aparecen debajo del contenido del mismo. El hecho de que se use una granularidad tan fina da una idea del valor que se da a la actualidad.

En las Figuras 6.1, 6.2 y 6.3 se muestran ejemplos de los elementos presentados.

6.1.2. Corpus de RepLab

RepLab [7] es un Foro de Evaluación de sistemas de gestión de reputación online.

Para la edición de 2013 se seleccionaron 60 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.

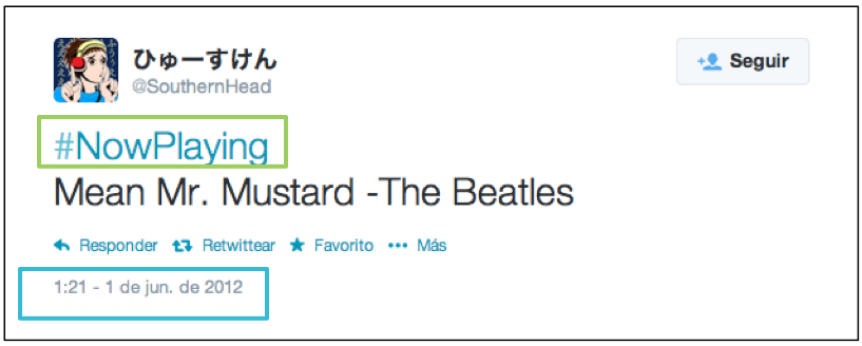


Figura 6.1: Ejemplo de *hashtag*

Figura 6.2: Ejemplos de mención y enlace

Figura 6.3: Ejemplo de *retweet*

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:

- *Relación o no con la entidad*
- *Polaridad*: se refiere a las implicaciones que pueda tener el contenido del *tweet* para la reputación de la entidad (positivo, negativo o neutro)
- *Tema tratado*
- *Prioridad*: da idea del grado de importancia que tiene la opinión sobre la entidad, en un determinado tema (relevante, medianamente relevante o irrelevante)

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. El conjunto de *tweets* restante, también se proporcionaba a los participantes. Éstos disponían, además, de información relativa a las páginas Web de las entidades, páginas de Wikipedia, etc.

6.1.3. Corpus Beatles

La entidad **"Beatles"** es elegida, de entre las 60, y sus *tweets* constituyen el corpus de experimentación: **"Corpus Beatles"**.

Composición

Está formado por un conjunto entrenamiento de 701 *tweets* (538 en inglés, 163 en español) y por uno de validación de 1531 *tweets* (1130 en inglés, 401 en español) .

Por cada *tweet* se dispone de la siguiente información: *identificador, autor, entidad, url, idioma* y *timestamp*. También contiene información relativa a las urls presentes en el contenido, así como un indicador de si es o no un *retweet*.

El texto del *tweet* se proporciona en archivo aparte, ya que las normas de Twitter no permiten el archivo y difusión de este dato, y ha de ser accedido directamente a través de su web; por esto, una parte importante de los *tweets* carece de contenido,

al haber desaparecido en este tiempo el mensaje o el autor. Concretamente, tras eliminar los *tweets* "vacíos", quedan 609 en el conjunto de entrenamiento y 1176 en el de validación.

Temporalización

El periodo temporal abarcado por ambos conjuntos es bastante pequeño y muestra una sorprendente regularidad.

El 98 % de los *tweets* de entrenamiento han sido publicados en una ventana de cinco días, mientras que el mismo porcentaje de los *tweets* de validación, que son casi el doble, lo han sido en una ventana de diez días (ver Cuadro 6.1).

ENTRENAMIENTO	VALIDACIÓN	
1 jun 2012: 69	22 dic 2012: 69	27 dic 2012: 87
2 jun 2012: 72	23 dic 2012: 77	28 dic 2012: 101
3 jun 2012: 85	24 dic 2012: 84	29 dic 2012: 112
4 jun 2012: 157	25 dic 2012: 71	30 dic 2012: 97
5 jun 2012: 76	26 dic 2012: 83	31 dic 2012: 86

Cuadro 6.1: Fechas de publicación Corpus Beatles

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 nuestros datos. Por ejemplo, se observa un incremento notable de actividad el día 4 de junio de 2012, fecha en que se produjo en Reino Unido el *Concierto del Jubileo*, en homenaje a la Reina.

Temática

Atendiendo a las anotaciones del conjunto de entrenamiento, se puede ver que en general los temas más habitualmente tratados están relacionados con comentarios de fans, letras y vídeos de canciones y referencias varias a productos (ediciones remasterizadas de discos, etc.) (ver Cuadro 6.2). 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*. Un ejemplo de esto sería el antes mencionado *concierto del Jubileo* o el *Album "SGT Pepper"* de

cuyo lanzamiento se cumplieron 45 años el 1 de junio de 2012; en este caso, el evento producido sería el aniversario del lanzamiento.

TEMAS GENÉRICOS	CUESTIONES ACTUALES
Comentarios sobre productos, comparativas, *merchandaising*... (26 %)	Concierto del Jubileo (5 %)
Comentarios de fans (19 %) y comentarios negativos (3 %)	SGT Pepper Album (3 %)
Letras (19 %) y vídeos (4 %)de canciones	Tributos, reediciones, remasterizaciones (3 %)

Cuadro 6.2: Conjunto de entrenamiento - Temática

Expresiones temporales

La anotación del Corpus se llevó a cabo con el sistema **HeidelTime** (ver Apartado 2.2.3). Dicho sistema anota los *tweets* en formato TimeML, marcando las expresiones temporales con la etiqueta TIMEX3 (ver Apartado 2.2.1).

Como se puede ver en el Cuadro 6.3, el porcentaje de *tweets* que contiene expresiones temporales no es muy grande, aunque sí significativo, roza el 16 %. Se puede ver también que hay *tweets* que contienen más de una expresión temporal, si bien no es la norma; hay que recordar que la extensión de los *tweets* es muy limitada.

	TWEETS	TWEETS CON TIMEX3	TIMEX3
ENTRENAMIENTO	609	104	132
VALIDACIÓN	1176	183	224

Cuadro 6.3: Expresiones temporales anotadas (HeidelTime)

Respecto al tipo de expresiones empleadas (ver Cuadro 6.4), la gran mayoría son de tipo DATE, esto es, representan fechas (en cualquiera de sus granularidades: año, mes, día). Notar también que hay un predominio de expresiones que denotan duración (DURATION), respecto a las que denotan momentos (TIME). Otro tipo de expresiones, relacionados con periodicidad, etc. (SET), son más bien raras.

	DATE	TIME	DURATION	SET
ENTRENAMIENTO	85	13	29	5
VALIDACIÓN	180	12	24	8

Cuadro 6.4: Tipología de las expresiones anotadas (HeidelTime)

Fechas

Las referencias a fechas son mayoritariamente en granularidades de años o días. Cuando se trata de días, suelen coincidir, o bien con la ***timestamp*** del tweet, provenientes de referencias relativas tipo *"today", "ayer"* o bien con **aniversarios**, esto es, con los mismos días pero de años pasados.

En el cuadro 6.5 se muestran las **expresiones normalizadas** de fecha, clasificadas por su referencia temporal al **pasado, presente** o **futuro**. *PRESENT_REF*, *PAST_REF* y *FUTURE_REF* son los valores con los que se normalizan expresiones indeterminadas como *"now", "recently", "próximamente", "no hace mucho"*, etc. Aquellas expresiones con una frecuencia inferior a 2 son descartadas para evitar incluir expresiones provenientes de errores o poco significativas.

Como se dijo antes, se observa un claro predomino de referencias a tiempo presente. Veamos unos ejemplos:

(1) <TIMEX3 tid="t4" type="DATE" value="PRESENT_REF">***Now***</TIMEX3> *playing on WildWest Radio: Ticket to Ride - by: The Beatles - From the Album: 1962-1966 (The Red Album) [2010 Remaster]*

(2) *Vamos a soñar imaginen q todos los integrantes d los beatles estuvieran vivos sería* <TIMEX3 tid="t1" type="DATE" value="2012-06-04">***hoy***</TIMEX3> *la locura en el concierto en honor a la reina*

En (1) la información temporal está relacionada con el momento de reproducción de una canción; es un tipo de referencia muy habitual, que tiene que ver con la forma en que se comparten experiencias o pensamientos en Twitter. En (2), por el contrario, se menciona un evento que se está celebrando, el *concierto del Jubileo* en honor a la Reina de Inglaterra.

Las referencias al pasado suelen aparecer de forma explícita, en *tweets* donde se data el contenido compartido (una canción, un vídeo) (3) o se menciona un evento

	PASADO	PRESENTE	FUTURO
ENTRENAMIENTO	1962 (3) 1964 (3) 1967 (9) 01/06/67 (4) 2009 (4)	PRESENT_REF (13) 2012-06-01 (9) 2012-06-02 (6) 2012-06-03 (3) 2012-06-04 (5)	
VALIDACIÓN	1962 (4) 1967 (4) 1969 (4) 2009 (4)	PRESENT_REF (36) 2012 (11) 22/12/2012 (4) 23/12/2012 (4) 24/12/2012 (6) 25/12/2012 (26) 27/12/2012 (4) 28/12/2012 (3) 29/12/2012 (5) 30/12/2012 (8) 31/12/2012 (3)	FUTURE_REF (4) 2013 (4)

Cuadro 6.5: Fechas de aparición más frecuente

de cierta relevancia, como la publicación de un álbum, un concierto, etc. (4).

(3) *Beatles Song Of The Day; Cant Buy Me Love Written at the GeorgeV hotel in Paris* <TIMEX3 tid="t1" type="DATE" value="1964">***1964***</TIMEX3> *by McCartney. Also recorded in Paris.*

(4) <TIMEX3 tid="t3" 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*

Por último, las referencias al futuro son escasas y en muchas ocasiones se debe a que los *tweets* están escritos en los últimos días del año:

5. *is going through another early Beatles listening phase :-) 1962-2012 Happy new* <TIMEX3 tid="t3" type="DATE" value="2013">***2013***</TIMEX3>. *Peace on Earth!*

Momentos

Las expresiones que refieren momentos son escasas y tienen que ver con referencias presentes pero concretas, esto es, en vez de provenir de adverbios como "ahora", lo hacen de expresiones como "esta noche", "this morning", etc, que son normalizadas teniendo en cuenta la *timestamp* del *tweet*. Así pues, aparecen también en *tweets* que contextualizan un comportamiento (6) o que anuncian algún tipo de evento en un momento temporal próximo (7).

(6) *starting* <TIMEX3 tid="t1" type="TIME" value="2012-06-02TMO">***this morning***</TIMEX3> *off right with the beatles and my bowl :) #empowering*

(7) *@USAndMumbai We will be recyling the beatles in our legendary hour at* <TIMEX3 tid="t1" type="TIME" value="XXXX-XX-XXT18:00">***6pm***</TIMEX3> *for starters .. :-)*

Duraciones

Respecto a las expresiones etiquetadas como duración (DURATION), suelen tener mucho que ver con fechas señaladas en la historia de la entidad, esto es, con aniversarios (8) o con periodos en los que se destaca algún aspecto de la entidad (9). En el caso de los Beatles, por supuesto, ha de referirse a una fecha de hace varias décadas; pero en todo caso, una duración inferior al año no aportaría información temporal significativa (10).

(8) *Hoy se cumplen* <TIMEX3 tid="t2" type="DURATION" value="P45Y">***45 años***</TIMEX3> *del estreno de Sgt. Pepper's Lonely Hearts Club Band, álbum de The Beatles.*

(9) *Los Beatles dominan las ventas de sencillos en* <TIMEX3 tid="t1" type="DURATION" value="P60Y">***los últimos 60 años***</TIMEX3>

(10) *Hoy* <TIMEX3 tid="t3" type="DURATION" value="P1D">***todo el día***</TIMEX3> x *música de The Beatles*

Set

Como ya se ha dicho, las expresiones de tipo SET son raras, si bien es cierto que cierto tipo de sintaxis se asocia a las formas de expresar datos estadísticos y puede resultar de interés (11).

(11) <TIMEX3 tid="t1" type="SET" value="P1Y" quant="EACH">***Each year*** </TIMEX3> *the Beatles continuously sell more records than the rolling stones*

Eventos

Se realiza también la anotación del subcorpus en inglés, con la herramienta **TARSQI-TTK** (ver apartado 2.2.3). Las anotaciones incluyen expresiones temporales (TIMEX3), eventos (EVENT), y relaciones entre eventos (LINK). No se consideran éstas últimas en este momento, y nos centramos exclusivamente en los eventos.

Al contrario de lo que ocurría para las expresiones temporales, el porcentaje de anotación es muy alto (ver Cuadro 6.6) y muchas veces se anotan varios eventos por *tweet*.

	TWEETS	TWEETS CON EVENTOS	EVENTOS	LEMAS
ENTRENAMIENTO	466	302	712	295
VALIDACIÓN	897	541	1153	351

Cuadro 6.6: Eventos anotados (TARSQI-TTK)

Las palabras etiquetadas como eventos son mayoritariamente verbos (12), aunque también hay sustantivos (13) o adjetivos (14). Los eventos de más frecuente aparición (después de un proceso de lematización) vienen reflejados en el Cuadro 6.7. La influencia del dominio se hace notar en la presencia de verbos como "listen" o "play".

(12) *This morning I have mostly been*<EVENT eid="e1" class="PERCEPTION">***listening***</EVENT> *to 'The Beatles - White Album'*

(13) *TONIGHT @thepeel *Beatles Tribute Band* Abbey Road LIVE! Sgt Pepper 45th Anniversary* <EVENT eid="e1" class="OCCURRENCE">***Show***</EVENT> *! $20! #avl #avlent #avlmusic #Asheville*

(14) *CHOON! The Beatles really were* <EVENT eid="e1" class="STATE" >***brilliant*** </EVENT> *! #jubileeconcert*

ENTRENAMIENTO		VALIDACIÓN	
do	24	get	43
get	20	listen	40
have	19	love	40
love	19	have	33
listen	17	be	32
play	17	do	29
make	16	put	28
think	14	stare	27
be	13	know	26
let	13	let	25

Cuadro 6.7: Eventos de más frecuente aparición

En cuanto al tipo de eventos, la inmensa mayoría son de ocurrencia (OCURRENCE) (13). También tienen una presencia significativa los de percepción (PERCEPTION) (12) y los de estado (ESTADO) (14). En el Cuadro 6.8 se muestra la distribución exacta de los tipos.

	ENTRENAMIENTO	VALIDACIÓN
ASPECTUAL	10	25
I_ACTION	6	11
I_STATE	46	69
OCCURRENCE	531	856
PERCEPTION	34	43
REPORTING	16	31
STATE	69	118

Cuadro 6.8: Tipos de eventos

De nuevo aprecia una notable regularidad entre el conjunto de entrenamiento y el de validación, tanto en los eventos que aparecen como en la distribución de los tipos.

6.2. Experimentos

Se parte de los *tweets* anotados en lenguaje TimeML. Para la colección de entrenamiento, se dispone también de la tabla goldstandard con datos del tema (*topic*) asignando manualmente. El resto de anotaciones no se van a considerar.

Como se dijo en el apartado anterior, casi todos los *tweets* se concentran en unos pocos días de Junio (conjunto de entrenamiento) y de Diciembre (conjunto de test). Dado que entre los *tweets* de fechas dispares existen algunos anteriores al 1 de Junio de 2012, fecha de referencia de la colección, entendemos que se han recopilado por error, por lo que se toma la decisión de no incluirlos en los experimentos. Al hacer esto, se puede suponer que la colección está compuesta por todos los *tweets* publicados en dichos días, y por tanto se pueden extraer conclusiones sobre el nivel de actividad de los usuarios.

Se elegirán distintos conjuntos de descriptores y se experimentará sobre dos conjuntos: el de entrenamiento, **BeatlesTra** y el total, **Beatles**. La batería completa de experimentos se resume en el Cuadro 6.9.

EXPERIMENTO	DESCRIPTORES
I	$\tau = \Phi$
II	$\tau = \Phi \bigcup T$
III	$\tau = \Phi \bigcup T \bigcup E$
IV	$\tau = \Phi \bigcup T \bigcup tipos(E)$
V	$\tau = tipos(E)$

Cuadro 6.9: Elección de descriptores

6.2.1. Experimento I

Se toman como descriptores $\tau = \Phi$, el conjunto de **fechas de creación** extendido de la siguiente manera: por cada *tweet* con fecha de creación *"YYYY-MM-DD"*, se toman los descriptores *"YYYY"*, *"YYYY-MM"* y *"YYYY-MM-DD"*.

Las características de los contextos formales generados se resumen en el Cuadro 6.10. Además, se muestra la representación de Hasse de ambos contextos (ver Figuras 6.4 y 6.5).

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesTra	693	7	7
Beatles	2195	18	19

Cuadro 6.10: Contexto formal Experimento I

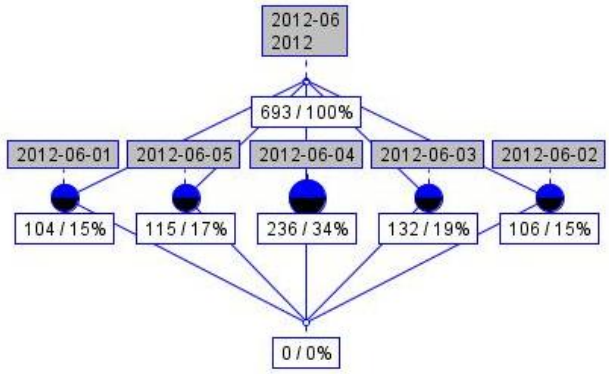


Figura 6.4: Diagrama de Hasse - BeatlesTra - ExpI

El retículo resultante tiene pocos conceptos formales, por lo que es sencillo. La elección de descriptores ha agrupado los *tweets* por día, mes y año; cada una de estas agrupaciones es un concepto formal.

El diagrama generado da idea del grado de actividad de los usuarios en relación a la entidad. No obstante, la ventana temporal que abarcamos es tan pequeña que no se pueden extraer muchas conclusiones. Si por ejemplo se contara con *tweets* de todo un año, una agrupación por meses o por semanas sí daría una idea de periodos de interés. Al menos deberíamos contar con los meses completos, ya que la razón de que en Diciembre haya prácticamente el doble de *tweets* es que la colección se ha confeccionado de esta manera.

Sí podemos observar que el día *4 de Junio*, la actividad fue notablemente más alta que 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 nuestro retículo no da esta información, pero nos advierte de que algún evento importante puede haber sucedido.

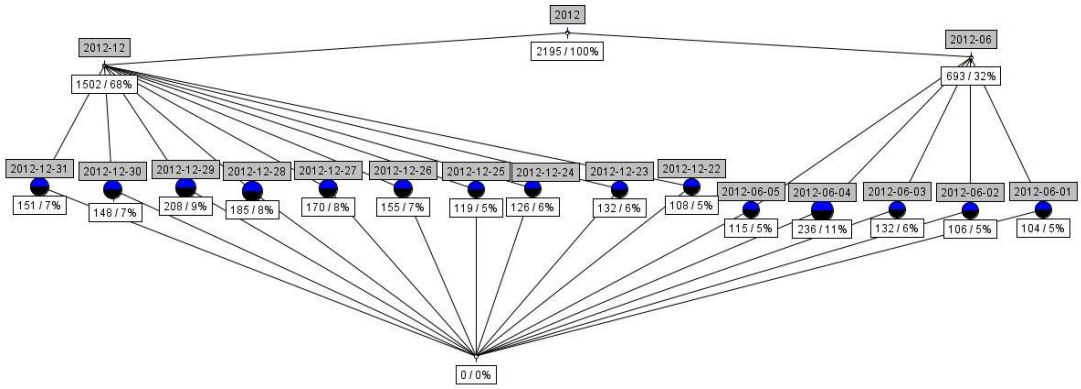


Figura 6.5: Diagrama de Hasse - Beatles - ExpI

6.2.2. Experimento II

Tomamos ahora $\tau = \Phi \bigcup T$.

Esto es, las **fechas de creación** y las **expresiones temporales** presentes en el contenido.

Respecto a las expresiones temporales, tomamos las siguientes decisiones:

- no tenemos en cuenta expresiones indeterminadas, de las identificadas con etiquetas PRESENT_REF, PAST_REF y FUTURE_REF, porque no compatibilizan bien con nuestro Calendario, según está construido.

- eliminamos las expresiones normalizadas de las palabras “Yesterday” y “Tomorrow” por provenir generalmente de canciones

- eliminamos las expresiones que solo aparecen una vez, por si se tratara de errores en la anotación

- extendemos cada expresión a las granularidades que refleja el Cuadro 6.11:

EXPRESION	EXTENSIÓN
YYYY-MM-DD	YYYY-MM YYYY MM-DD
YYYY-MM	YYYY

Cuadro 6.11: Descriptores asociados a cada expresión

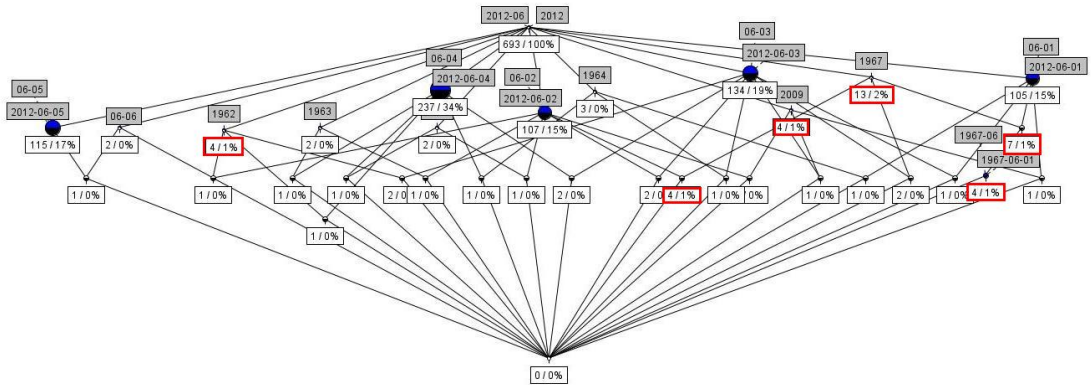


Figura 6.6: Diagrama de Hasse - BeatlesTra - ExpII

- extendemos también las fechas de creación a la granularidad *MM-DD*[2]

Los retículos de conceptos generados se describen en el Cuadro 6.12. Observamos que el número de conceptos supera significativamente al número de atributos (especialmente en el conjunto **BeatlesTra**), lo que quiere decir que se han encontrado relaciones entre los objetos.

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesTra	693	21	35
Beatles	2195	42	47

Cuadro 6.12: Contexto formal Experimento II

Si nos fijamos en el retículo generado para **BeatlesTra** (Figura 6.6), podemos ver que muchos de los nuevos conceptos contienen un único objeto. Estos conceptos marcan apariciones de fechas muy poco frecuentes. También vemos que se crean varios conceptos con 4 o más objetos (marcados en rojo en la figura). Estas agrupaciones son las que nos interesan porque buscamos organizar los *tweets* según su contenido temporal.

Puesto que contamos con las anotaciones de la *gold-standard* y con un pequeño "truco", representamos el mismo diagrama, pero mostrando el tema de los *tweets* en lugar de su identificador (ver Cuadro 6.7).

[2]Esta extensión no aportaba información nueva en el Experimento I y por eso no fue incluida

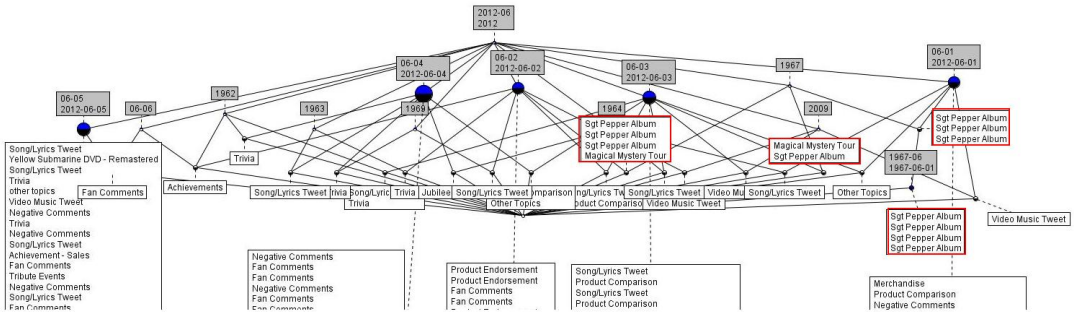


Figura 6.7: Distribución de temas - BeatlesTra - ExpII

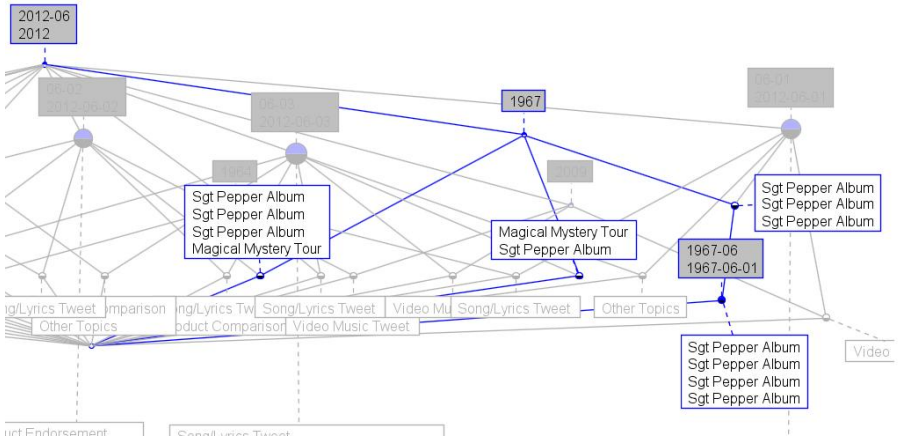


Figura 6.8: *Tema SgtPepper*

Encontramos que FCA ha aislado aceptablemente parte del tema *"SgtPepper"*. 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 (un 3 % de los *tweets* trataban el tema, como vimos el estudiar el Corpus). En la figura 6.8 vemos en detalle el concepto en que se basa la agrupación. Como se ve, prácticamente todos los *tweets* que hacen referencia al año 1967, lo hacen al álbum SgtPepper; y todos los *tweets* escritos el 1 de junio que hacen referencia a 1967, también.

En cuanto al conjunto completo, observamos que no se crean relaciones en Diciembre. Extrañamente, aunque el conjunto de *tweets* tenía el doble de tamaño que el

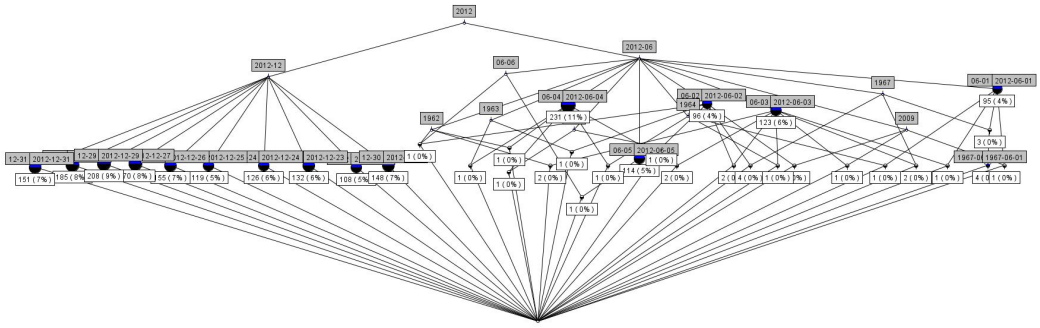


Figura 6.9: Diagrama de Hasse - Beatles - ExpII

de entrenamiento, no se encontró presencia de expresiones temporales con frecuencia mayor que 2, por lo que las que aparecieron fueron descartadas. La percepción es, que a nivel temporal, los temas de los *tweets* están separados por la brecha en la fecha de creación.

6.2.3. Experimento III

Buscamos ahora nuevas relaciones completando el conjunto de descriptores con los **eventos** más frecuentes anotados por Tarsqi, $\tau = \Phi \bigcup T \bigcup E$. Concretamente imponemos que los eventos estén presentes en, al menos, un 3 % de los *tweets*, lo que supone unos 60 *tokens*.

El número de conceptos se dispara (ver Cuadro 6.13), y también el de relaciones, perdiendo totalmente el control sobre el retículo. Los conceptos más poblados corresponden a *tweets* cuyos eventos no han sido considerados descriptores y por tanto están completamente descritos por su fecha de creación (ver Figura 6.10). Además el nivel de agrupamiento conseguido para el tema "Sgt.Pepper" se pierde.

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesTra	297	75	339
Beatles	830	108	996

Cuadro 6.13: Contexto formal Experimento III

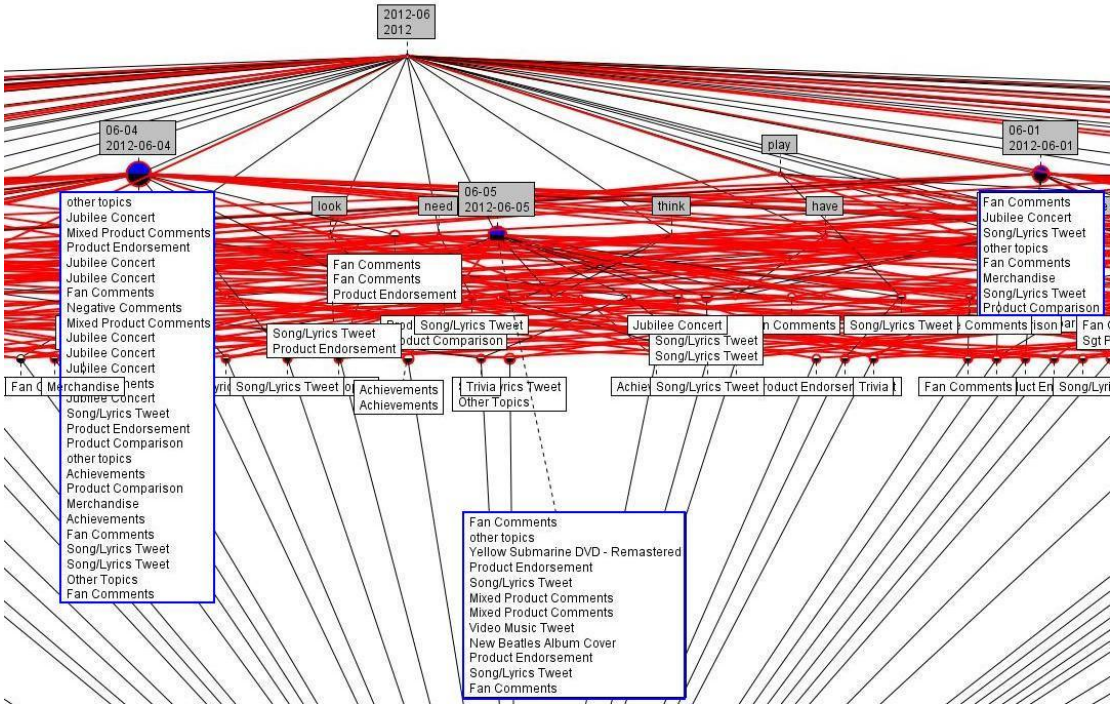


Figura 6.10: Detalle Diagrama de Hasse - BeatlesTra - ExpIII

6.2.4. Experimento IV

En vista de los malos resultados del Experimento III, vamos a tratar de aprovechar la información sobre los eventos de los *tweets* utilizando, en vez del evento, su **tipo**: $\tau = \Phi \bigcup T \bigcup tipos(E)$. Esto permitirá reducir significativamente el número de atributos, resumiendo los9 aproximadamente 60 *tokens* de eventos en sólo siete, uno por cada tipo de evento (ocurrencia, percepción, estado,...).

Pese a aumentar solo en 7 atributos los contextos del Experimento II, el número de conceptos creados es muy grande (ver Cuadro 6.14). Esto se debe a que la densidad de esos atributos es muy alta en los *tweets*.

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesTra	297	26	115
Beatles	830	47	297

Cuadro 6.14: Contexto formal Experimento IV

6.2.5. Experimento V

Por último decidimos explorar la tipología de eventos como único argumento, $\tau = tipos(E)$.

El número de conceptos obtenido es bastante menor, como se ve en el Cuadro 6.15.

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesTra	297	7	30
Beatles	830	7	36

Cuadro 6.15: Contexto formal Experimento V

La primera conclusión que podemos extraer es en relación a los predominancia de los eventos de ocurrencia 6.11. Por otro lado, aunque parece que los conceptos cuya intensión es un único tipo (por ejemplo, *tweets* con eventos de percepción) están llamados a congregar un cierto tipo de *tweet*, la forma en que lo hacen es muy genérica. Veamos por ejemplo, los eventos de tipo “estado” del retículo del conjunto

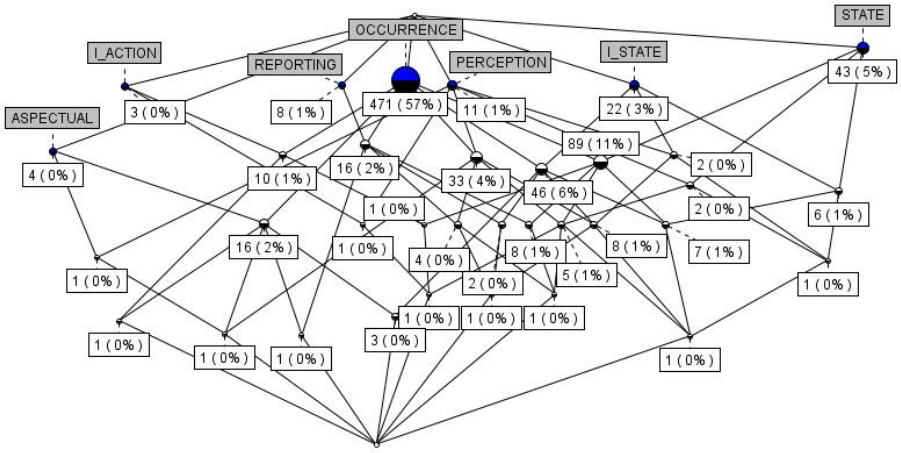


Figura 6.11: Diagrama de Hasse - Beatles - ExpV

de entrenamiento (ver Figura 6.12). Parecen ser *tweets* que expresan opinión, pero lo hacen tanto para hablar de un producto, como de una canción,etc.

6.3. Valoración de resultados

Las posibilidades de evaluación de nuestra propuesta, teniendo en cuenta que no hemos considerado la colección completa y dadas las características de ésta, son reducidas.

Nuestro primer objetivo era responder a la pregunta: *¿cómo se expresa la información temporal en Twitter?* En la descripción del Corpus hemos detallado el tipo de expresiones que aparecen y las hemos relacionado con ciertos temas; por ejemplo hemos visto que es frecuente "datar" un contenido que se comparte.

El segundo objetivo pasaba por organizar temporalmente los *tweets*. Los resultados del Experimento I prueban que FCA permite una representación informativa del grado de actividad en Twitter relacionado con una entidad en un determinado periodo de tiempo. Para extraer conclusiones es básico conocer los datos que se están examinando, en concreto, si se dispone de todos los *tweets* correspondientes a ese

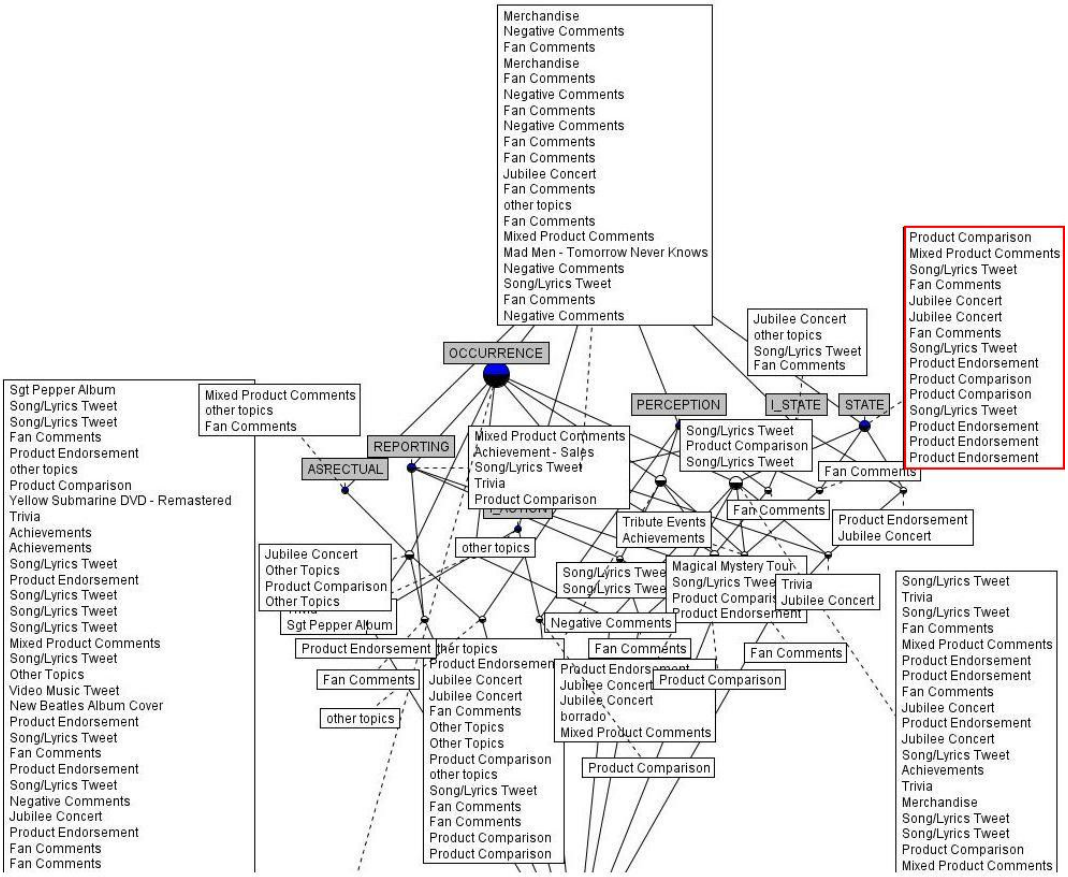


Figura 6.12: Detalle Diagrama de Hasse - BeatlesTra - ExpV

periodo.

El tercer objetivo iba un paso más allá, buscando organizar los *tweets* por los eventos tratados, bien por la fecha en que se produjeron, bien por el evento en sí, bien por el tipo de evento. Los mejores resultados se produjeron utilizando un conjunto de descriptores basado en fechas de creación y expresiones temporales (Experimento II). Obviamente, y dado que la presencia de dichas expresiones es significativa pero reducida en Twitter, hay un limitado tipo de temas que podemos aspirar a detectar. Se aisló con éxito el tema "Sgt.Pepper", sin embargo, queremos hacer una precisión: el tema realmente detectado por nuestro retículo fue "publicación del disco Sgt.Pepper", y cualquier otro *tweet* que hubiera hecho referencia a dicho álbum por cualquier otra razón, no habría sido relacionado. Estamos aquí ante la ya conocida dualidad "evento *vs* topic".

Teniendo en cuenta esto último tenemos que decir que, aún habiendo hecho el estudio para toda la colección, y de habernos sometido al sistema de evaluación de RepLab, esperaríamos los mejores resultados para temas identificables con eventos, pero en nuestro afán de contextualizar temporalmente los *tweets* perderíamos efectividad en temas de carácter genérico.

Finalmente, hemos de comentar que los retículos generados, que buscan maximizar la relación entre el momento de creación del *tweet* y su contenido, requieren del manejo simultáneo de múltiples granularidades. Esto da, en general, un contexto con una serie de atributos con un alto grado de redundancia, a la vez que una gran frecuencia de aparición, que son los correspondientes a las fechas de creación. Estas peculiares características son las responsables de que no hayamos podido explotar todas las capacidades del entorno Cigarrán et al. [16]; pues sus algoritmos de reducción tendían a resumir el retículo eliminando las relaciones generadas con las expresiones temporales extraidas del contenido.

6.4. Pruebas y experimentación con otros Corpus

Con el objetivo de averiguar si los resultados obtenidos para el Corpus serían extrapolables a otros conjuntos de datos, se repite la mejor configuración del experimento (Experimento II) para los siguientes conjuntos:

- **BeatlesBackground**. Formado por 50000 *tweets* sobre los Beatles, es por tanto un conjunto veinte veces mayor que la colección de prueba.
- **Bankia**. Conjuntos de entrenamiento y test de la entidad "Bankia". Es una entidad bancaria y por tanto de un dominio muy diferente al de la colección de prueba; además se trata de una entidad española, por lo que la mayor parte de los *tweets* proceden de España.

Se toma pues $\tau = \Phi \bigcup T$ y se pretende saber si el modelo propuesto es capaz de detectar y aislar temáticamente eventos de marcada naturaleza temporal, esto es, que se identifiquen con una fecha.

6.4.1. BeatlesBackground

El contexto formal resultante tiene las características que se resumen en el Cuadro 6.16.

	OBJETOS	ATRIBUTOS	CONCEPTOS
BeatlesBackground	5647	51	89

Cuadro 6.16: Contexto formal BeatlesBackground (Experimento II)

Se forman, al menos, dos conceptos formales que aislan conjuntos de *tweets* relacionados directamente con fechas. Aunque no disponemos de la tabla *gold-standard*, puesto que conocemos la fecha, podemos saber a qué dos eventos se refieren. El primero de ellos se muestra en la Figura 6.13, y se trata de la fecha en que John Lennon consiguió la Green Card (permiso para residir en EEUU). El segundo de ellos (Figura 6.14), se relaciona con el "5 de octubre de 1962", fecha en que los Beatles lanzaron la canción *"Love me Do"*. En este último caso, el tamaño del concepto es notable, esto es, agrupa un gran número de *tweets* (75), máxime teniendo en cuenta que todos han sido publicados el mismo día (5 de octubre de 2012). Lo cierto es que en dicha fecha, se cumplían 50 años de la citada canción y las celebraciones, felicitaciones, etc. lo llevaron a ser *"trending topic"*[3].

[3]http://www.aztecatrends.com/notas/musica/130302/Love-Me-Do-de-The-Beatles-cumple-50-aos

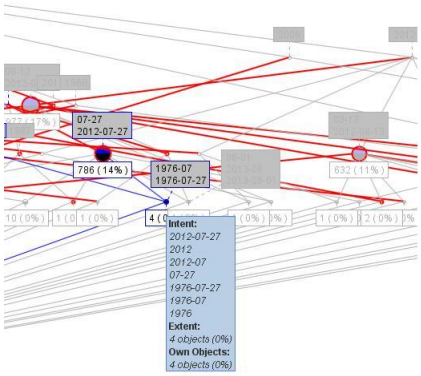


Figura 6.13: Detalle Diagrama de Hasse - BeatlesBackground - *1976-07-27*

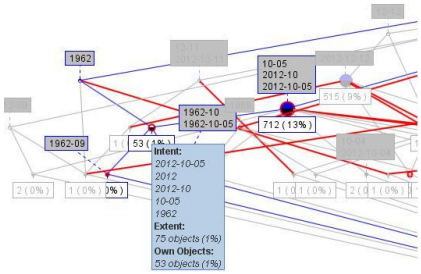


Figura 6.14: Detalle Diagrama de Hasse - BeatlesBackground - *1962-10-05*

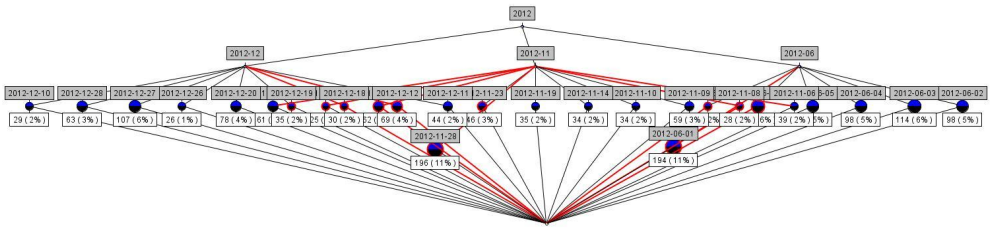


Figura 6.15: Diagrama de Hasse - Bankia

6.4.2. Bankia

El contexto formal resultante se describe en el Cuadro 6.15.

	OBJETOS	ATRIBUTOS	CONCEPTOS
Bankia	1839	73	88

Cuadro 6.17: Contexto formal Bankia (Experimento II)

Lo primero que se observa es que hay un gran número de atributos, en relación a la entidad Beatles. Esto es debido, en parte, a que la relevancia de la entidad es de carácter nacional (aunque algunos temas hayan sido *"trending topic"* global) y por tanto, la ventana necesaria para completar los conjuntos de entrenamiento y test es considerablemente mayor. Además hay bastante variabilidad en el número de *tweets* publicado por día; destacan especialmente el "1 de junio" y el "18 de noviembre de 2012" (ver Figura 6.15). Ambas fechas están relacionadas con noticias de gran impacto en el país, como fueron, la ocupación de varias sedes por los llamados "yayoflautas"[4] y el anuncio del E.R.E[5].

Se forma un concepto formal que contiene dos *tweets* relacionados con una noticia publicada el "9 de febrero de 2012" (ver Figura 6.16). El contenido de esos *tweets*, publicados en Junio, es una crítica al contenido de la noticia y por extensión a la entidad, y como tal se anota en la tabla *gold-standard*. En este punto, volvemos a incidir en que la representación temporal no es capaz de agrupar *tweets* en temas

[4]http://www.20minutos.es/noticia/1490225/0/yayoflautas/madrid-barcelona-valencia-sevilla-palma/bankia/

[5]http://www.elmundo.es/elmundo/2012/11/28/economia/1354096837.html

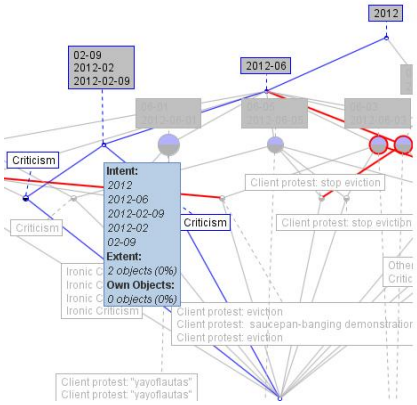


Figura 6.16: Detalle Diagrama de Hasse - Bankia - *2012-02-09*

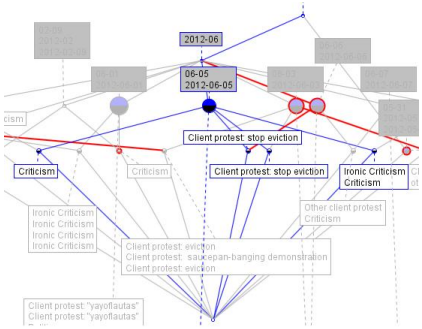


Figura 6.17: Detalle Diagrama de Hasse - Bankia - *2012-06-05*

como *"críticas"*, sino que encuentra los eventos concretos que sirven de vehículo a éstas.

Por otro lado, el concepto formal del "5 de junio" tiene varios subconceptos relacionados con las protestas contra los desahucios (ver Figura 6.17). Se trata de convocatorias concretas de actuaciones ciudadanas para impedir la acción de la policía, etc. De nuevo cada actuación irá a parar a un concepto diferente. Se da la circunstancia de que las convocatorias que se producen en el mismo día de su celebración, se confunden con el grueso de *tweets* publicados, por lo que en este aspecto, el modelo es mejorable.

Parte III

Conclusiones finales

Capítulo 7

Conclusiones y líneas futuras de investigación

A lo largo de este trabajo se ha tratado de poner en valor la información temporal asociada a los documentos, para distintas tareas de Recuperación de Información.

Las principales contribuciones se resumen en los siguientes puntos:

1. Análisis de corpus de *tweets* bajo la perspectiva temporal.
2. Propuesta de un modelo para la representación de la información temporal de *tweets*, basado en Análisis Formal de Conceptos.
3. Adaptación e integración de recursos web y paquetes software para experimentación.
4. Experimentación y valoración de resultados en un sub-corpus de la colección de RepLab 2013.

Concretamente se ha intentado crear un marco temporal en el que contextualizar conjuntos de *tweets*, relacionando el evento referido con la fecha de creación y expandiendo esas relaciones para buscar similitudes entre los *tweets*. El **Análisis Formal de Conceptos** se adapta naturalmente al modelo construido: los conceptos formales son las agrupaciones de *tweets* que tratan un evento, los *tweets* son los objetos y los atributos, aquéllos que tratan de definir el evento en base a su fecha de creación, las

expresiones temporales que contiene, etc. El carácter temporal del retículo está derivado de la representación multigranular de la información; cada expresión temporal y cada fecha de creación, se representan en múltiples granularidades, de forma que pasar de una a otra supone una especialización o generalización del concepto.

La **anotación** automática, al igual que cualquier otro Procesamiento del Lenguaje que hagamos en Twitter presenta una problemática derivada de la idiosincrasia del dominio. El lenguaje de anotación estándar, TimeML, nació para la anotación temporal de corpus de noticias; dichos documentos poseen un estilo característico en el que son comunes ciertos elementos, como el establecimiento de un contexto de referencia o el volver sobre un tema que ya ha sido tratado. En Twitter el contexto está "fuera" del *tweet*, los comentarios de otros usuarios, la persona que escribe, las características de su perfil, el momento exacto en que el *tweet* se publica,... constituyen una información muy valiosa para la interpretación de su contenido. En nuestra propuesta nos centramos en el **contexto temporal**, para detectar cierto tipo de *tweets* que se escriben *porque el momento es el que es*. Con la definición del **Calendario Imaginario-Colectivo** quisimos capturar ese conocimiento.

Al contrario que en los textos de noticias, en los *tweets* no abundan las expresiones temporales. En el estudio del Corpus hemos hecho un análisis del tipo de expresiones que suelen aparecer y con qué tipo de *tweets* se identifican. Una línea de trabajo futuro en este campo, podría ser la **anotación de hashtags con información temporal**, del estilo del famoso #15m.

Respecto a las anotaciones de las que dispusimos, dejamos sin explorar las relaciones entre eventos, etiquetadas con la **etiqueta *LINK***. Sería interesante ver si su explotación ayuda en la selección de los descriptores, ya que los resultados de los experimentos en los que utilizamos eventos fueron malos y no daban la sensación de estar representando adecuadamente el contenido del *tweet*. La tipología de eventos, dio resultados más interesantes, sobre todo fuera del amplio conjunto de eventos de "ocurrencia". Otros tipos de información que no se utilizaron, fueron las referencias indeterminadas al presente, pasado o futuro así como a los momentos de duración inferior al día. A nuestro modo de ver su capacidad para identificar temas es menor que la del resto de expresiones, y agregan complejidad al contexto.

La representación de la información mediante Análisis Formal de Conceptos

mostró capacidad para aislar eventos que se correspondían exactamente con fechas, pero hay otros tipos de análisis que podrían ser objeto de estudio posterior, como la **duración de los temas**, medida en función de las distintas fechas de creación de los *tweets* que los tratan, etc.

Por último, sería conveniente definir alguna medida de **evaluación** de la calidad de los retículos generados.

Bibliografía

[1] James Allan, Rahul Gupta, y Vikas Khandelwal. Temporal summaries of new topics. En *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, págs. 10–18. ACM, 2001.

[2] James F Allen. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843, 1983.

[3] James F Allen. Towards a general theory of action and time. *Artificial intelligence*, 23(2):123–154, 1984.

[4] Omar Alonso, Ricardo Baeza-Yates, y Michael Gertz. Effectiveness of temporal snippets. En *WSSP Workshop at the World Wide Web Conference WWW*, tomo 9. 2009.

[5] Omar Alonso, Michael Gertz, y Ricardo Baeza-Yates. Clustering and exploring search results using timeline constructions. En *Proceedings of the 18th ACM conference on Information and knowledge management*, págs. 97–106. ACM, 2009.

[6] Omar Alonso, Jannik Strötgen, Ricardo A Baeza-Yates, y Michael Gertz. Temporal information retrieval: Challenges and opportunities. *TWAW*, 11:1–8, 2011.

[7] Enrique Amigó, Jorge Carrillo de Albornoz, Irina Chugur, Adolfo Corujo, Julio Gonzalo, Tamara Martín-Wanton, Edgar Meij, Maarten de Rijke, y Damiano Spina. Overview of replab 2013: Evaluating online reputation monitoring systems. En Pamela Forner, Henning Müller, Roberto Paredes, Paolo Rosso, y Benno Stein, eds., *CLEF*, tomo 8138 de *Lecture Notes in Computer Science*, págs. 333–352. Springer, 2013. ISBN 978-3-642-40801-4.

[8] Sitaram Asur, Bernardo A Huberman, Gabor Szabo, y Chunyan Wang. Trends in social media: Persistence and decay. En *ICWSM*. 2011.

[9] Ricardo A. Baeza-Yates, Julien Masanès, y Marc Spaniol, eds. *Proceedings of the 1st International Temporal Web Analytics Workshop (TWAW 2011), Hyderabad, India, March 28, 2011*, tomo 707 de *CEUR Workshop Proceedings*. CEUR-WS.org, 2011. URL http://dblp.uni-trier.de/db/conf/www/twaw2011.html.

[10] Steven Bird. Nltk: the natural language toolkit. En *Proceedings of the COLING/ACL on Interactive presentation sessions*, COLING-ACL '06, págs. 69–72. Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1225403.1225421. URL http://dx.doi.org/10.3115/1225403.1225421. Última visita: 2014-05-27.

[11] Ricardo Campos, Gaël Dias, Alípio Mário Jorge, y Célia Nunes. Enriching temporal query understanding through date identification: how to tag implicit temporal queries? En *Proceedings of the 2nd Temporal Web Analytics Workshop*, págs. 41–48. ACM, 2012.

[12] Claudio Carpineto y Giovanni Romano. Effective reformulation of boolean queries with concept lattices. En *Flexible Query Answering Systems*, págs. 83–94. Springer, 1998.

[13] Angel Castellanos, Juan Cigarrán, y Ana García-Serrano. Modelling techniques for twitter contents: A step beyond classification based approaches. En *Working Notes of the CLEF 2013*. 2013. URL http://www.clef-initiative.eu/documents/71612/86f6c2c1-5148-4c18-a41e-a59a0d9f0fa1.

[14] Nancy A. Chinchor. Proceedings of the Seventh Message Understanding Conference (MUC-7) named entity task definition. En *Proceedings of the Seventh Message Understanding Conference (MUC-7)*, pág. 21 pages. Fairfax, VA. URL http://acl.ldc.upenn.edu/muc7/ne_task.html. Version 3.5.

[15] Juan Manuel Cigarrán Recuero. *Organización de resultados de búsqueda me-*

diante análisis formal de conceptos. Tesis Doctoral, Universidad Nacional de Educación a Distancia, 2008.

[16] Juan Cigarrán, Angel Castellanos, y Ana García-Serrano. Herramienta de análisis de contenido basada en AFC. 2014. 1ª versión.

[17] Gaël Harry Dias, Mohammed Hasanuzzaman, Stéphane Ferrari, y Yann Mathet. Tempowordnet for sentence time tagging. En *Proceedings of the Companion Publication of the 23rd International Conference on World Wide Web Companion*, WWW Companion '14, págs. 833–838. 2014. ISBN 978-1-4503-2745-9. URL http://dx.doi.org/10.1145/2567948.2579042. Última visita: 2014-05-15.

[18] Lisa Ferro, Laurie Gerber, Inderjeet Mani, Beth Sundheim, y George Wilson. Tides 2005 standard for the annotation of temporal expressions. 2005.

[19] Lisa Ferro, Inderjeet Mani, Beth Sundheim, y George Wilson. Tides temporal annotation guidelines-version 1.0. 2. *The MITRE Corporation, McLean-VG-USA*, 2001.

[20] Elena Filatova y Eduard Hovy. Assigning time-stamps to event-clauses. En *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, pág. 13. Association for Computational Linguistics, 2001.

[21] Antony Galton. A critical examination of allen's theory of action and time. *Artif. Intell.*, (2-3):159–188. URL http://dblp.uni-trier.de/db/journals/ai/ai42.html#Galton90. Última visita: 2014-04-01.

[22] Bernhard Ganter y Rudolf Wille. *Formal Concept Analysis: Mathematical Foundations*. Springer, Berlin/Heidelberg, 1999.

[23] Guillermo Garrido, Anselmo Penas, Bernardo Cabaleiro, y Alvaro Rodrigo. Temporally anchored relation extraction. En *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, págs. 107–116. Association for Computational Linguistics, 2012.

[24] Robert Godin, C Pichet, y J Gecsei. Design of a browsing interface for information retrieval. En *ACM SIGIR Forum*, tomo 23, págs. 32–39. ACM, 1989.

[25] Iqbal A Goralwalla, Yuri Leontiev, M Tamer Özsu, Duane Szafron, y Carlo Combi. Temporal granularity: Completing the puzzle. *Journal of Intelligent Information Systems*, 16(1):41–63, 2001.

[26] Bernard J Jansen, Mimi Zhang, Kate Sobel, y Abdur Chowdury. Twitter power: Tweets as electronic word of mouth. *Journal of the American society for information science and technology*, 60(11):2169–2188, 2009.

[27] Haewoon Kwak, Changhyun Lee, Hosung Park, y Sue Moon. What is twitter, a social network or a news media? En *Proceedings of the 19th international conference on World wide web*, págs. 591–600. ACM, 2010.

[28] Juha Makkonen, Helena Ahonen-Myka, y Marko Salmenkivi. Topic detection and tracking with spatio-temporal evidence. En *Advances in information retrieval*, págs. 251–265. Springer, 2003.

[29] Inderjeet Mani y George Wilson. Robust temporal processing of news. En *Proceedings of the 38th Annual Meeting on Association for Computational Linguistics*, págs. 69–76. Association for Computational Linguistics, 2000.

[30] Drew McDermott. A temporal logic for reasoning about processes and plans. *Cognitive Science: A Multidisciplinary Journal*, págs. 101 – 155. ISSN 0364-0213. URL http://www.informaworld.com/10.1207/s15516709cog0602_1. Última visita: 2014-04-01.

[31] D. Odijk, G. Santucci, M. de Rijke, M. Angelini, y G. Granato. Time-aware exploratory search: Exploring word meaning through time. En *SIGIR 2012 Workshop on Time-aware Information Access*. Portland, OR, USA, 2012.

[32] United States. Defense Advanced Research Projects Agency. Information Technology Office. *Sixth Message Understanding Conference, (MUC-6): Proceedings of a Conference Held in Columbia, Maryland, November 6-8, 1995*. ISBN 9781558604025. URL http://books.google.es/books?id=n3RQAAAAMAAJ. Última visita: 2014-04-11.

[33] J Pustejovsky. Terqas: time and event recognition for question answering systems (2002).

[34] James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, y Dragomir R Radev. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34, 2003.

[35] James Pustejovsky, Patrick Hanks, Roser Sauri, Andrew See, Robert Gaizauskas, Andrea Setzer, Dragomir Radev, Beth Sundheim, David Day, Lisa Ferro, et al. The timebank corpus. En *Corpus linguistics*, tomo 2003, pág. 40. 2003.

[36] James Pustejovsky, Bob Ingria, Roser Sauri, Jose Castano, Jessica Littman, Rob Gaizauskas, Andrea Setzer, Graham Katz, y Inderjeet Mani. The specification language timeml. *The language of time: A reader*, págs. 545–557, 2005.

[37] Ridho Reinanda, Daan Odijk, y Maarten de Rijke. Exploring entity associations over time.

[38] Estela Saquete y Patricio Martínez-Barco. Grammar specification for the recognition of temporal expressions. *Proceedings of Machine Translation and multilingual applications in the new millennium. MT2000, Exeter, UK*, págs. 21–1, 2000.

[39] Estela Saquete, Patricio Martinez-Barco, y Rafael Muñoz. Recognizing and tagging temporal expressions in spanish. En *Workshop on Annotation Standards for Temporal Information in Natural Language, LREC*, tomo 2002, págs. 44–51. Citeseer, 2002.

[40] Estela Saquete, Rafael Munoz, y Patricio Martínez-Barco. Event ordering using terseo system. *Data & Knowledge Engineering*, 58(1):70–89, 2006.

[41] R Saurı, O Batiukova, y J Pustejovsky. Annotating events in spanish. timeml annotation guidelines (version tempeval-2010). Inf. téc., Barcelona Media. Technical Report BM 2009-01, 2009.

[42] Roser Saurí, Estela Saquete, y James Pustejovsky. Annotating time expressions in spanish. *TimeML Annotation Guidelines. Version TempEval-2010*, 2010.

[43] Frank Schilder y Christopher Habel. From temporal expressions to temporal information: Semantic tagging of news messages. En *Proceedings of the workshop on Temporal and spatial information processing-Volume 13*, pág. 9. Association for Computational Linguistics, 2001.

[44] Helmut Schmid. Treetagger— a language independent part-of-speech tagger. *Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart*, 43, 1995.

[45] Andrea Setzer. *Temporal Information in Newswire Articles: An Annotation Scheme and Corpus Study*. Tesis Doctoral, University of Sheffield, 2001.

[46] Andrea Setzer y Robert J Gaizauskas. Annotating events and temporal information in newswire texts. En *LREC*, tomo 2000, págs. 1287–1294. 2000.

[47] Jannik Strötgen y Michael Gertz. Heideltime: High quality rule-based extraction and normalization of temporal expressions. En *Proceedings of the 5th International Workshop on Semantic Evaluation*, págs. 321–324. Association for Computational Linguistics, Uppsala, Sweden, 2010. URL http://www.aclweb.org/anthology/S10-1071.

[48] Naushad UzZaman, Hector Llorens, James Allen, Leon Derczynski, Marc Verhagen, y James Pustejovsky. Tempeval-3: Evaluating events, time expressions, and temporal relations. *arXiv preprint arXiv:1206.5333*, 2012.

[49] Marc Verhagen, Robert Gaizauskas, Frank Schilder, Mark Hepple, Graham Katz, y James Pustejovsky. Semeval-2007 task 15: Tempeval temporal relation identification. En *Proceedings of the 4th International Workshop on Semantic Evaluations*, págs. 75–80. Association for Computational Linguistics, 2007.

[50] Marc Verhagen, Inderjeet Mani, Roser Sauri, Robert Knippen, Seok Bae Jang, Jessica Littman, Anna Rumshisky, John Phillips, y James Pustejovsky. Automating temporal annotation with tarsqi. En *Proceedings of the ACL 2005 on Interactive poster and demonstration sessions*, págs. 81–84. Association for Computational Linguistics, 2005.

[51] Marc Verhagen y James Pustejovsky. The tarsqi toolkit. En Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan,

Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, y Stelios Piperidis, eds., *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*. European Language Resources Association (ELRA), Istanbul, Turkey, 2012. ISBN 978-2-9517408-7-7.

[52] María Teresa Vicente-Díez, César de Pablo-Sánchez, y Paloma Martínez. Evaluación de un sistema de reconocimiento y normalización de expresiones temporales en español. 2007.

[53] María Teresa Vicente-Diez y Paloma Martinez. Temporal semantics extraction for improving web search. En *Database and Expert Systems Application, 2009. DEXA'09. 20th International Workshop on*, págs. 69–73. IEEE, 2009.

[54] Wikipedia. Iso 8601. . URL http://es.wikipedia.org/wiki/ISO_8601#Historia_del_est.C3.A1ndar. Última visita: 2014-04-11.

[55] Wikipedia. Tiempo unix. . URL http://es.wikipedia.org/wiki/Tiempo_Unix. Última visita: 2014-05-26.

[56] Serhiy Yevtushenko, Julian Tane, Tim B. Kaiser, Sergei Obiedkov, Joachim Hereth, y Heiko Reppe. Conexp - the concept explorer. 2000-2006. URL http://conexp.sourceforge.net. Última visita: 2014-05-21.

[57] Arkaitz Zubiaga, Damiano Spina, Enrique Amigó, y Julio Gonzalo. Towards real-time summarization of scheduled events from twitter streams. En *Proceedings of the 23rd ACM conference on Hypertext and social media*, págs. 319–320. ACM, 2012.