Extracción y asignación de tiempos de suceso para documentos de actualidad

Dolores Llidó Escrivá, Rafael Berlanga Llavori y Mª José Aramburu Cabo

Departament d'Informática
Campus Riu Sec, Universitat Jaume I
E-12003 Castellón, ESPAÑA
e-mail: { dllido, berlanga, aramburu}@nuvol.uji.es
TF: 9647283 {60,67,04}

En este trabajo se presenta una aproximación a la problemática de determinar el tiempo de suceso de un documento estructurado. Esta propuesta se basa en la extracción de expresiones temporales a partir de las sentencias lingüísticas que aparecen en los textos, y que se refieren al tiempo. Estas expresiones temporales servirán para determinar el tiempo de suceso de cada frase, párrafo y finalmente de todo el documento. Por último, indicaremos el impacto que puede tener el tiempo de suceso en la recuperación de los documentos con las técnicas clásicas de los Sistemas de Recuperación de la Información.

1. Introducción

Una buena parte de los documentos electrónicos que manejamos a diario nos informan acerca de sucesos y tópicos localizados en el tiempo. Un claro ejemplo de éstos son las noticias, los informes médicos, etc. Este tipo de documentos plantean nuevos desafíos en el área de la Recuperación de la Información debido principalmente a dos razones:

- Estos documentos requieren nuevos modelos de datos y consulta para expresar la temporalidad de los documentos [1].
- Dado que la información se localiza en el tiempo, resulta interesante ver como evoluciona el espacio de la información. Concretamente, las tareas de detectar nuevos sucesos y monitorizar tópicos en los repositorios son dos temas de investigación de reciente aparición [2].

Las técnicas clásicas de Recuperación de la Información no resultan adecuadas para abordar este tipo de aplicaciones [3]. Por un lado, los modelos de documentos están muy orientados a los términos textuales, y en menor medida a la estructura de los documentos. La temporalidad de estos documentos no puede expresarse en estos modelos y los lenguajes de recuperación sólo permiten trabajar con atributos como la fecha de publicación.

En [1] se presenta una aproximación basada en las bases de datos temporales orientadas a objeto. El resultado es un modelo de base de datos, denominado TOODOR, el cual permite representar cualquier tipo de documento estructurado junto con sus propiedades temporales. Además, el lenguaje de consulta de este modelo permite recuperar los documentos por su contenido, estructura y tiempo. Sin embargo, esta propuesta parte de la hipótesis de que cada documento debe tener un periodo de suceso asignado de forma manual por un experto. Claramente, esta hipótesis no puede asumirse en aplicaciones donde el flujo de documentos es enorme, como por ejemplo sucede en la Web.

En este artículo planteamos una aproximación a la obtención automática de estos periodos de suceso. La idea general es analizar el contenido de los documentos para identificar las sentencias temporales de los textos e interpretarlas semánticamente a partir de un modelo de tiempo. El resultado de este análisis será un etiquetado del texto con las fechas extraídas. En un proceso posterior, estas fechas se utilizan para determinar el periodo de suceso de los documentos.

2. Modelos Semánticos

En esta sección describiremos los modelos semánticos necesarios para procesar los documentos de nuestras aplicaciones. En primer lugar mostraremos brevemente el modelo de representación de los documentos, y a continuación el modelo de tiempo que permitirá determinar las referencias temporales de los textos.

2.1. Modelo de representación de documentos

Para representar los documentos hemos utilizado el modelo TOODOR [1], el cual no describiremos en detalle pues excede el ámbito de este artículo.

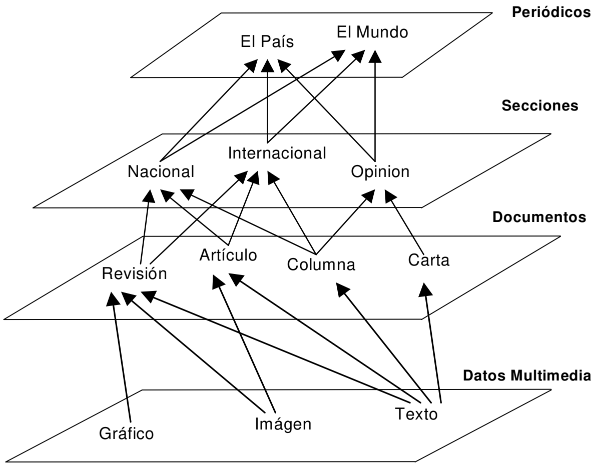

Figura 1. Ejemplo de jerarquía de agregación de documentos en el modelo TOODOR.

Brevemente, los documentos según este modelo se representan como objetos XML [4] con identidad propia y con dos atributos temporales: el tiempo de inserción y el periodo de suceso. El primer atributo juega un papel importante en la extracción de referencias temporales ya que sirve de punto de referencia para determinar ciertas expresiones lingüísticas tales como “hoy”, “mañana”, etc. En cuanto al periodo de suceso, éste se considera el intervalo máximo que cubre los acontecimientos de actualidad descritos en el documento. Este atributo es el que permite recuperar la información de los documentos con mayor precisión, así como monitorizar y detectar los sucesos descritos en la base documental. En este artículo trataremos de extraer este atributo de forma automática a partir de los textos.

El modelo TOODOR además permite componer documentos complejos a partir de otros más simples mediante el mecanismo de agregación del modelo orientado a objeto. Por ejemplo, un periódico puede componerse de secciones que a su vez se componen de artículos. De esta forma los documentos estructurados se representan como árboles cuyas hojas contienen la información textual y multimedia. Esta estructura es necesaria para determinar los periodos de suceso de los documentos, ya que éstos se extraen de los textos y se propagan hacia arriba en la jerarquía de agregación.

2.2. Modelo de tiempo

En lenguaje natural, la forma de construir sentencias sobre el tiempo se basa en el uso de calendarios, los cuales se basan a su vez en las denominadas granularidades de tiempo. A partir de éstas se pueden construir sentencias lingüísticas para expresar instantes e intervalos de tiempo, duraciones, etc. En este orden describiremos los elementos del modelo de tiempo subyacente a las sentencias temporales.

2.2.1. Granularidades

La base del modelo de tiempo es el sistema de granularidades mostrado en la figura 1. De ahora en adelante, denotaremos cada granularidad con una letra minúscula: día (d), semana (w), quincena (q), mes (m), cuatrimestre (q), semestre (s), año (y), década (c) y siglo (x). Estas granularidades se organizan según una relación de refinamiento, denotada $\prec$ [3], la cual se muestra en la figura con flechas.

Cada granularidad g puede verse como un tipo de datos, donde el dominio del tipo, denotado $dom(g)$, describe los posibles valores que podemos manejar en cada granularidad (ej. $dom(m)=\{1..12\}$). En este trabajo, todas las granularidades tendrán definido su dominio sobre los números naturales. A lo largo del artículo utilizaremos las funciones $first(g)$ y $last(g)$ para denotar el primer y último elemento del dominio de g.

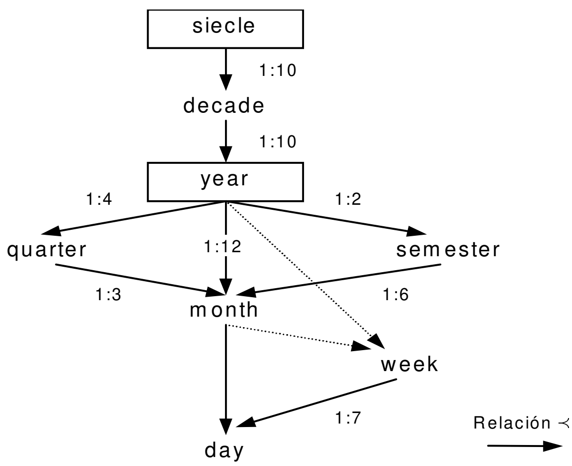

Figura 2: Sistema de granularidades

Es importante mencionar que las granularidades recuadradas en la figura son aquellas cuyo dominio es absoluto, es decir, cada valor de su dominio representa un intervalo determinado de fechas. Como veremos más adelante, para determinar una expresión temporal se necesitará tener al menos una de estas granularidades definida.

La conversión de valores entre granularidades se realiza aplicando un factor de escala, denotado *ratio*(g, g'), que en la mayor parte de los casos es fijo (ver figura 2). Sin embargo, ciertas conversiones (flechas discontinuas) no podrán realizarse aplicando un factor fijo, sino que hará falta consultar el calendario correspondiente. Este es el caso de la conversión de meses a días, de semanas a días, de años a semanas y de meses a semanas. La mayor parte de los modelos de tiempo de la literatura no soportan este tipo de conversiones, que por otro lado suelen ser usuales en lenguaje natural (ej. "la primera semana del 2000").

2.2.2. Entidades temporales

A partir del sistema de granularidades ya podemos construir las entidades temporales. En primer lugar expresaremos un *punto de tiempo* como una secuencia alternada de granularidades y valores:

$$T = g_1\, n_1\; g_2\; n_2 \ldots g_k\, n_k$$

donde $n_i \in dom(g_i)$ para todo $1 \leq i \leq k$, y las granularidades están ordenadas de mayor a menor ($g_{i+1} \prec g_i$).

Un punto de tiempo denota un instante expresado con un determinado grado de imprecisión, el cual viene dado por la granularidad más fina. Denotaremos la granularidad más fina de un punto T con $gran(T)$.

Un caso especial de punto de tiempo son las *fechas* del calendario, cuyo formato es "$\mathtt{y}\ \mathtt{n}_1\ \mathtt{m}\ \mathtt{n}_2\ \mathtt{d}\ \mathtt{n}_3$"

Un *intervalo de tiempo* es el espacio temporal entre dos instantes de tiempo. Al igual que los puntos de tiempo, los intervalos también pueden expresarse con distintos grados de imprecisión mediante el uso de granularidades. Así, un intervalo se expresará en el modelo como un par de puntos de tiempo con la misma granularidad:

$$I = [T_1,\ T_2]$$

donde $T_1 = g_1\, n_1 \ldots g_k\, n_k$, $T_2 = g_1\, n'_1 \ldots g_k\, n'_k$
y $n_i \leq n'_i$ para todo $1 \leq i \leq k$

Al igual que para los puntos, definiremos la granularidad de un intervalo como $gran(I) = gran(T_1) = gran(T_2)$. Además, denotaremos con $start(I)$ y $end(I)$, el principio y final del intervalo I respectivamente.

Finalmente definimos una *duración* de tiempo como una extensión temporal dirigida hacia el pasado o el futuro. Esta entidad también puede expresarse con distintas granularidades para definir distintos grados de precisión. Seguiremos la siguiente notación para las duraciones:

$$S = \pm\, n_1\, g_1 \ldots n_k\, g_k$$

donde el signo indica la dirección de la extensión (+ futuro y – pasado), los valores n_i ($1 \leq i \leq k$) son número naturales, y g_i ($1 \leq i \leq k$) son las granularidades ordenadas de mayor a menor ($g_{i+1} \prec g_i$).

Algunos ejemplos de duraciones son los siguientes: `-1m` , `+1y 2m`, `+2y3m2d`.

2.2.3. Operaciones sobre entidades temporales

En este apartado describiremos las operaciones más importantes para la extracción de referencias temporales de los textos.

- *Refinamiento de entidades temporales:*

El refinamiento de un punto temporal a una granularidad inferior siempre produce un intervalo temporal, el cual expresa la imprecisión del punto temporal. Así, si tenemos un punto $T = n_1\, g_1\;\; n_2\, g_2 \ldots n_k\;\; g_k$ y $g \prec g_k$ entonces

$$refine(T, g) = [T_1,\ T_2]$$

donde $T_1 = g_1\, n_1\, g_2\, n_2 \ldots g_k\, n_k\; g\;\; first(g)$,
y $T_2 = g_1\, n_1\; g_2\, n_2 \ldots g_k\, n_k\; g\; last(g)$

De modo similar, podemos definir el refinamiento de un intervalo:

$$refine(I,g)= [start(refine(start(I),g)),end(refine(end(I),\ g))]$$

Veamos varios ejemplos:

refine(y1999,m)= [y1999m1, y1999m12]
refine(y2000m3,w)=[y2000m3w1,y2000m3w5]
refine([y2000,y2001],m)=[y2000m1,y2001m12]

- *Abstracción de un punto temporal:*

Esta operación es la inversa de la anterior, y permite abstraer puntos e intervalos a granularidades superiores. Es importante mencionar que el proceso de abstracción supone una pérdida de información ya que se pierden los detalles de las granularidades inferiores. Denotaremos la abstracción de un punto temporal $T = n_1\, g_1 \quad n_2\, g_2 \ldots n_k \quad g_k$ a una granularidad g como:

$$abstract(T, g)$$

Una propiedad interesante de esta operación es que puede dar como resultado un punto o un intervalo temporal dependiendo del tipo de conversión entre la granularidad mínima del punto (intervalo) y la granularidad objetivo. Veamos estos dos casos:

abstract(y2000m3d1, y) = y2000
abstract(y2000m1d1, w) = [y1999m12w59, y2000m1w1]

- *Desplazamientos de puntos temporales:*

Mediante los desplazamientos podemos obtener nuevos puntos e intervalos aplicando duraciones de tiempo con signo. Definiremos esta operación como sigue:

$$shift(T, S) = refine(S, gran(T)) + T$$

Aquí el operador + es la suma aritmética sobre los valores del dominio de la granularidad mínima de las entidades. Esta suma debe tener en cuenta el posible acarreo sobre las granularidades anteriores. La operación de desplazamiento solo está definida cuando $gran(T) \prec= gran(S)$.

A continuación mostramos algunos ejemplos de esta operación:

shift(y1999m3, +10m) = y2000m1
shift(y1998m2w2, -3w) = y1998m1w4

3. Extracción de referencias temporales

El objetivo de este trabajo es la extracción de referencias temporales a partir del análisis de los textos de los documentos de la aplicación. En este artículo nos concentraremos en la extracción de fechas en el formato del calendario gregoriano "$yn_1\, mn_2\, dn_3$".

Las sentencias del texto de las que se extraen las fechas serán etiquetadas en el documento original para su posterior análisis en la asignación de los periodos de suceso. Así, un ejemplo de este etiquetado sería el siguiente:

```
El Gobierno británico está
decidido a impedir que la marcha
de los unionistas de la Orden de
Orange prevista para el <TIMEX
TYPE='DATE' VALUE=19990627>
domingo </TIMEX>
```

Para este etiquetado seguiremos las directivas del MUC-7 [5], donde se define la etiqueta de tiempo `TIMEX`. En este formalismo, las expresiones temporales se clasifican como"`TIME`" si tienen una granularidad inferior a día, o "`DATE`" en el otro caso. Nosotros además hemos definido un nuevo atributo, `VALUE`, para guardar el valor de la fecha extraída del texto.

A grandes rasgos, los principales problemas que se plantean para la detección de expresiones temporales y la extracción de una fecha concreta a partir de una expresión son:

- El reconocimiento de las palabras que forman una misma expresión temporal. En un principio pensamos que este problema se podría resolver con un análisis sintáctico, pero probando algunas frases con la herramienta de análisis morfo-sintáctico de frases en español TANCAT [8] comprobamos que no se extraían automáticamente las expresiones temporales, se requería otro proceso para unir distintos sintagmas y así obtener la expresión temporal. Este proceso es muy complicado y requiere un gran tiempo de procesado, además de poseer un buen diccionario. Está claro que el tiempo de procesado se puede disminuir analizando solo aquellas sentencias que sabemos contienen expresiones temporales. A partir del estudio de las expresiones temporales hemos observado que estas siguen bien unos patrones fijos o bien que estas poseen un núcleo y unos modificadores que determinan la expresión. Además, en lenguajes como el castellano podemos encontrar frases largas, donde el orden de las palabras puede variar sin alterar su significado temporal (ej. dos días después, después de dos días), aunque si tienen distintas connotaciones semánticas en la frase. Por ello el algoritmo propuesto trata de detectar bien unos patrones o bien unas palabras clave (ej. día, semana, julio, semestre, martes...) en las sentencias del documento.

- Pasar de la expresión temporal a un punto o intervalo fijo de nuestro calendario o bien reconocer que se trata solo de una duración no anclada en el tiempo.

- Para fijar la fecha a la que se refiere la expresión temporal muchas veces no es suficiente con detectar la expresión temporal, sino que también se requiere el tiempo verbal. (ej. hace 2 semanas)
- Fijación de las indeterminaciones cuantitativas que se producen en el lenguaje natural. Los valores de las granularidades pueden aparecer indefinidos, bien porque se utilizan adjetivos indefinidos "algunos días después", o bien porque simplemente se utiliza el plural, "al final del día", "el final de los días".
- Resolución de coreferencias (ej. durante los dos días siguientes...) y deixis (ej. hoy, ayer..).

La aproximación que proponemos en este trabajo consiste en aplicar una mezcla de análisis sintáctico-semántico [6]. La idea general es identificar una serie de palabras clave, denominados *núcleos*, para detectar las sentencias temporales mediante expresiones regulares. A partir de éstas se detectan y traducen los términos lingüísticos a elementos del modelo de tiempo. Con ello, se obtiene una expresión temporal según el modelo que debe ser resuelta, para obtener finalmente las referencias temporales. En la siguiente sección se describen los elementos gramaticales necesarios para este proceso, y en las siguientes secciones se describen por separado cada una de las etapas del proceso.

En la figura 3 se muestra el proceso de extracción de fechas y etiquetado de los documentos de entrada. Estas etapas se discutirán en detalle en las siguientes secciones.

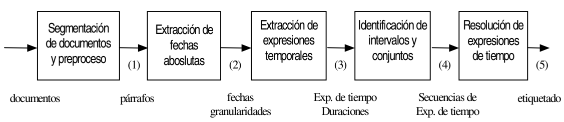

Figura 3: Proceso de extracción de referencias temporales

3.1. Elementos gramaticales

A partir del estudio de los sintagmas temporales en lenguaje natural hemos clasificado las palabras que los componen en tres categorías, las cuales nos van a proporcionar la información semántica necesaria para extraer las referencias temporales. Estas son:

1. *Núcleos*: son palabras relacionadas con las granularidades temporales. En concreto, representan las propias granularidades (ej. día, mes, etc.), sus sinónimos (ej. jornada), así como las palabras que denotan los valores de las granularidades (lunes, julio, etc.)
2. *Cuantificadores*: llamaremos así a todos los adjetivos cardinales, ordinales e indefinidos.
3. *Modificadores*: todas las formas gramaticales involucradas en una sentencia que hable acerca del tiempo. Estas pueden clasificarse a su vez en:
   - artículos, preposiciones, etc., que aunque no tienen una asociación directa con entidades del modelo temporal, permiten relacionar los distintos elementos que conforman dichas entidades.
   - palabras que expresan periodos o intervalos como “durante”, “entre”, etc.
   - palabras que indican la dirección temporal, como los adverbios de tiempo (ej. después, antes, etc.) y algunas formas verbales muy usuales (ej. hace, duró, etc.)
   - palabras que indican una parte de un intervalo temporal (ej. principio, final, mediados, mitad, etc.).

3.2.1. Segmentación del documento

Los documentos de entrada del proceso están estructurados en párrafos. Para segmentar estos párrafos en frases, hemos adoptado la aproximación presentada en [7]. Para identificar sentencias más pequeñas dentro de estas frases, aplicamos el mismo algoritmo, pero utilizando como separadores de sentencias los paréntesis, comas y guiones.

Por otro lado, en cada sentencia se reconocen las expresiones numéricas que contienen puntos, comas o signos- , así como los adjetivos cuantificadores traduciéndolos al formato del modelo de tiempo. Adicionalmente, los cuantificadores indefinidos se codifican de forma especial con el símbolo ‘x’.

A partir de este momento, las sentencias temporales se reescriben en cada etapa con los elementos identificados y codificados.

Fecha de publicación 3-07-99				
El 4 de mayo de 1999	y1999m05d05			
En mayo de este año	En m05 y1999	En y1999m05	[y1999m05]	
El próximo día 15	próximo d15	El 15+		19990715
principios el año pasado	principios y pasado	principios –1y	[-1y]	[1998]=refine(-1y,y)+1999
Hace un año	Hace 1 y	–1y	-1y	19980703=refine(-1y,d)+19990703
El próximo mes de mayo	El próximo m05	El m05+		200005
Los próximos días 14,15,16	Los próximos d14,d15,d16	(d14,d15,d16)+		(19990714,19990715,199900716)
Los próximos días	Los próximos ds	Los +ds	Los [+3d]	[19990703,19990706]=[19990703, refine(+3d,d)+19990703]
La primera semana de abril de este año	La w1 m04 y1999	La 1999m04w1		19990401,19990407]=[first(199904), refine(1w,d)+first(199904))-1]
Durante los 2 días anteriores a	Durante 2ds anteriores a	Durante -2ds a	[-r2d]	[T, *refine*(*-2d, gran*(*T*))+T]
Desde el martes hasta el jueves	Desde d102 hasta d104		[d102,d104]	[19990706-19990708]

Figura 4: Evolución del procesado de las expresiones temporales

3.2.2. Identificación de granularidades

El primer objetivo de esta etapa es extraer las expresiones temporales absolutas detectando bien días especiales o bien ciertos patrones sintácticos. Estas expresiones se traducen inmediatamente al formato del modelo (ej. “4 de marzo de 1999” → “`y1999m3d4`”).

Para simplificar el proceso posterior de extracción de expresiones temporales codificaremos además todas aquellas granularidades que aparecen en el texto (ej. “el junio pasado” → “el `m6` pasado”). Para esta tarea haremos uso del diccionario de núcleos, donde cada palabra tiene asociado su codificación para el modelo.

3.2.3. Identificación de las expresiones temporales relativas

En esta etapa se determinan los puntos temporales que pueden fijarse con la fecha de publicación (ej. “hoy”, “el próximo lunes”, etc.)

Además, se identifican las duraciones de tiempo (ej. "durante 2 días", "el día anterior", etc.) según la notación del modelo.

Para realizar estas tareas, se aplica un patrón sintáctico que busca alrededor de cada granularidad identificada un modificador que lo ancle en el tiempo (ej. “anterior”) y un cuantificador (ej. “dos semanas antes”).

3.2.4. Identificación de intervalos o listas

En esta fase el algoritmo intenta agrupar las granularidades codificadas, con el fin de reconocer puntos, intervalos, listas y duraciones temporales. Para realizar este agrupamiento nos basaremos en la distancia en el texto de las granularidades. Si esta distancia es mayor de cuatro palabras en el texto codificado asumiremos que se trata de dos entidades temporales distintas. Si la distancia es menor de dos y son granularidades distintas, entonces formarán una entidad temporal que se traducirá al modelo de tiempo (ej. “en `m5 y1999`” → `y1999m5`). Para el resto de casos, se aplicarán patrones sintácticos para reconocer intervalos ó listas de entidades temporales (ej. “desde `m2` hasta `m4`”→ “`[m2,m4]`”)

3.2.5. Obtención de fechas

Una vez hemos obtenido las entidades temporales a partir del texto, la última etapa se ocupa de operar sobre ellas para obtener fechas concretas.

Las entidades temporales extraídas pueden dividirse en dos categorías:

- *entidades exactas*: son aquellas que poseen referencias temporales explícitas o dependientes de la fecha de publicación.

- *entidades ambiguas*: son aquellas que dependen de otras referencias citadas en el texto (ej. "tres días después de la firma"), o bien puntos temporales no anclados en el tiempo (ej. "el martes").

La resolución de expresiones temporales ambiguas requiere un análisis semántico del texto, tema en el cual se está comenzando a investigar y que requiere un gran coste computacional [8]. Por ello, hemos optado por la utilización de suposiciones que en la mayoría de caso suelen ser ciertas, aunque pueden dar lugar a posibles errores en la extracción de la fecha. Estas suposiciones deben modificarse según el tipo de documentos de estudio. En el caso de periódicos se ha observado que se produce poco error si se toma siempre la fecha más próxima a la citada anteriormente en el texto.

Para concluir esta sección, presentamos en la figura 4 varios ejemplos de procesamiento según las etapas anteriores.

4. Obtención de periodos de evento

En la sección anterior hemos visto como extraer las referencias temporales de los textos. En esta sección veremos cómo se distribuyen en el tiempo y cómo puede asignarse el periodo de suceso a partir de dicha distribución.

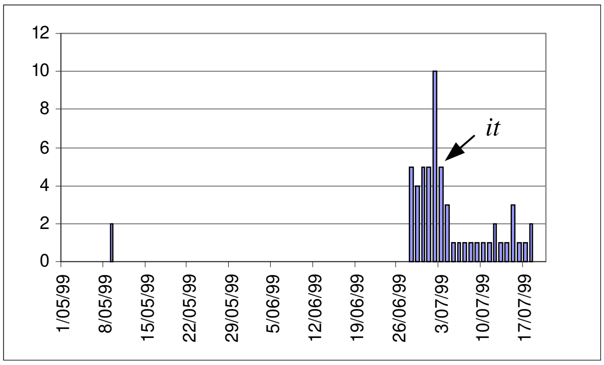

Figura 5: Histograma temporal de un documento de actualidad

En primer lugar, partiremos de un histograma típico de la distribución de las referencias temporales en un documento de actualidad, en este caso una noticia de un periódico. Cada barra del histograma representa una fecha extraída del texto mediante el método propuesto, y su altura el número de ocurrencias de dicha fecha en el documento.

Generalmente, el máximo del histograma está en el día anterior a la publicación del documento (*it*). Alrededor del máximo también hallamos numerosas referencias, que será las de interés para calcular el periodo de suceso del documento. Por otro lado, mucho más alejadas del máximo, podemos encontrar otras referencias temporales, que probablemente corresponda con sucesos distintos, aunque relacionados con los descritos por el documento.

El problema de determinar el periodo de suceso de un documento consiste en discernir qué referencias corresponden a sucesos relacionados y cuáles al propio suceso del documento. Para ello, la distancia a la fecha del máximo es un factor muy importante, pero también lo es la relevancia de cada fecha, pues esta podría indicarnos si lo descrito en el documento es una continuación de un suceso muy importante acontencido con anterioridad.

Para la aplicación periodística, hemos considerado las *referencias históricas* como aquellas que están alejadas más de dos meses de la fecha de publicación. Por otro lado, las fechas alejadas más de una semana de la fecha de publicación deben tener más de una ocurrencia para incluirlas en el periodo de suceso.

En la tabla 1 mostramos los primeros resultados obtenidos después de analizar cerca de 800 frases temporales de un periódico digital entero. Esta tabla clasifica los documentos según su distribución de referencias temporales:

- Clase A: documentos que solo contienen la fecha de publicación. Se trata de documentos breves que describen un suceso aislado sin relación con otros anteriores.
- Clase B: son documentos que contienen la fecha de publicación y una referencia histórica.
- Clase C: documentos que contienen varias referencias históricas y el periodo de suceso abarca varios días. Estos son documentos que describen hechos y sucesos importantes con numerosas relaciones a otros sucesos anteriores.
- Clase D: documentos que contienen muchas referencias históricas distribuidas uniformemente en el tiempo. Generalmente son cronologías sobre algún tema. La asignación de un periodo de suceso a estos documentos no tiene mucho sentido.
- Clase E: documentos que contienen un periodo de suceso extenso pero no hay referencias históricas. Estos suelen aparecen en las secciones de economía y deportes, donde los sucesos redactados

suelen tener una validez de semanas o meses.

A	B	C	D	E
17%	30,5%	23,2%	17,1%	12,2%

Tabla 1: Clasificación de documentos según su histograma temporal.

Como puede apreciarse, este tipo de documentos presentan una gran variedad de distribuciones temporales, y por tanto de periodos de suceso. Además, podemos constatar el alto número de referencias históricas que contienen estos documentos (alrededor del 70% de los documentos analizados). Estos factores nos indican por un lado que el periodo de suceso es un atributo necesario para recuperar con precisión parte de estos documentos, y que el gran número de referencias históricas en los textos puede degradar la precisión y *recall* de los sistemas de recuperación clásicos.

5. Conclusiones

En este artículo hemos presentado una aplicación para la extracción de fechas a partir de documentos de actualidad. Hemos obtenido un método que nos permite analizar las expresiones temporales sin requerir un análisis sintáctico completo de todo el documento.

A partir de las fechas del texto se puede obtener el periodo de suceso de cada documento, lo que facilitará el desarrollo de sistemas de recuperación de mayor precisión. A este respecto, hemos observado que en muchos casos el periodo de suceso de un artículo no siempre incluye la fecha de publicación, lo que hace necesario definir nuevos modelos de recuperación de la información que tengan en cuenta estos periodos de suceso, así como las referencias históricas que puedan aparecen en los documentos.

Agradecimientos

Este trabajo ha sido subvencionado por el proyecto CICYT con número de contrato TEL97-1119, y la Fundació Bancaixa de Castelló.

Referencias

[1] M. J. Aramburu and R. Berlanga "TOODOR: Un modelo de bases de datos para bibliotecas digitales" Novática n.142, pp. 10-15, December, 1999.

[2] J. Allan, R. Papka and V. Lavrenko "On-Line New Event Detection and Tracking". SIGIR, pp. 37-45, 1998.

[3] M.J.Aramburu and R.Berlanga "A Retrieval Language for Historical Documents" 9th International Conference on Database and Expert System Applications, LNCS 1460, pp. 216-225, Springer Verlag, 1998.

[4] D.Llidó, M.J. Aramburu, R. Berlanga y I. Sanz "Representación y organización de periódicos digitales con el lenguaje XML" IV Congreso ISKO-España, pp. 171-178, Granada, Abril 1999.

[5] Named Entity Definition según la Versión 3.5 (1997 - Nancy Chinchor). URL: http://www.muc.saic.com/proceedings/ne_Task.html)

[6] R. Grishman." Information Extraction: Tehniques and Challenges. International Summer School" SCIE-97. Ed. Maria Teresa Pazienza, Springer-Verlag, pp 10-27, 1997.

[7] J. Peral, P. Martínez-Barco, R. Muñoz, A. Ferrández, L. Moreno, M. Palomar, "Una técnica de análisis parcial sobre textos no restringidos (SUPP) aplicada a un Sistema de Extracción de Información (EXIT)". VI Simposio Internacional de Comunicación Social. Cuba, 1999

[8] J. Atserias, J. Carmona, I. Castellón, S. Cervell, M. Civit, L. Màrquez, M.A. Martí, L. Padró, R. Placer, H. Rodríguez, M. Taulé y J. Turmo "Morphosyntactic Analysis and Parsing of Unrestricted Spanish Text". First International Conference on Language Resources and Evaluation (LREC'98). Granada, Spain, 1998.