UNIVERSITAT POLITÈCNICA DE VALÈNCIA

DIPLOMA DE ESPECIALIZACIÓN EN BIG DATA

PROYECTO FIN DE DIPLOMA

Estudio de la evolución temporal de temas basado coocurrencias y cohesión léxica

Curso 2014-2015
1ª Edición

Autor
Isabel Edo

Director
Francisco Manuel Rangel Pardo

Índice

1. Introducción

El análisis de redes sociales, como una fuente de información del entorno social y competitivo de las empresas, es una tarea que ha cobrado gran relevancia en los sistemas de vigilancia tecnológica e inteligencia competitiva de las organizaciones.

En este trabajo se aborda una de las técnicas de análisis de redes sociales, basada en el análisis del texto de los contenidos de los mensajes publicados y concretamente en el análisis de los términos y sus relaciones. Se utilizará la coocurrencia de palabras como medida de similitud semántica, que indica que si dos términos tienden a aparecer juntos, es que comparten un significado. A estas relaciones de similitud se les denomina *relatedness*, ya que no indican el tipo exacto de relación existente pero que se adquiere en el contexto del análisis. Este análisis nos permitirá conocer mejor el contenido de los mensajes, identificando las relaciones que establecen los términos de nuestro interés con otros términos, permitiendo identificar mejor de qué se está hablando sobre nuestra marca, producto o sector. Este análisis se complementa con el análisis de cohesión léxica de los textos para identificar la densidad o dispersión léxica entre el conjunto de los textos.

Desde siempre, las empresas, de una u otra forma, han estado pendientes del progreso tecnológico que se genera en el entorno mediante la lectura de revistas técnicas, la asistencia a ferias de muestras, el examen de los productos de los competidores, el análisis de las patentes, etc. Se recopila información del entorno económico, político y social y se analiza para obtener conocimiento que ayude a la toma de decisiones.

La Vigilancia Tecnológica se define como “sistema organizado de observación y análisis del entorno, tratamiento y circulación interna de los hechos observados y posterior utilización en la empresa” (Palop y Vicente, 1999).

La norma UNE 166006 se define como “el proceso organizado, selectivo y sistemático, para captar información del exterior y de la propia organización sobre ciencia y tecnología, seleccionarla, analizarla, difundirla y comunicarla, para convertirla en conocimiento con el fin de tomar decisiones con menor riesgo y poder anticiparse a los cambios”.

La Inteligencia Competitiva es definida por Gibbons y Prescott (1996): Inteligencia Competitiva es el proceso de obtención, análisis, interpretación y difusión de información de valor estratégico sobre la industria y los competidores, que se transmite a los responsables de la toma de decisiones en el momento oportuno.

Muchas veces se confunden ambos conceptos al referirse a la misma cosa. Es obvio que los dos están estrechamente relacionados, sobre todo, porque la inteligencia competitiva es una evolución de la Vigilancia. Por ello, no son fácilmente distinguibles.

La Vigilancia nace ante la necesidad de las empresas de observar su entorno y así poder responder a determinados cambios cuando éstos se producen. La Inteligencia Competitiva, sin embargo, parte del conocimiento del entorno, lo cual implica poder adelantarse a los cambios, entendidos en ambos casos, como las amenazas y las oportunidades.

De esta manera, la diferencia fundamental entre las dos disciplinas es la actitud activa de la Inteligencia, para "conocer" el entorno y así anticiparse a los cambios. Ello supone no esperar a ver dónde se producen los cambios para actuar después, sino buscar activamente las oportunidades del entorno, lo que conlleva una revolución en la manera de entender todas las actividades gerenciales, comerciales y de innovación de la empresa.

La vigilancia tecnológica y la inteligencia competitiva se orientan a proporcionar información para la toma de decisiones, pero su objeto de análisis marca la diferencia entre ellas. Mientras la primera se enfoca en el seguimiento de la evolución de la tecnología y de sus implicaciones, la segunda lo hace en otros factores de competitividad, como los competidores actuales y potenciales, clientes, proveedores, entorno normativo, etc., y sus repercusiones en la competitividad de las empresas.

Las redes sociales son una fuente de información de la Inteligencia competitiva. Algunos autores la incluyen dentro de la llamada "inteligencia de fuentes abiertas", OSINT en inglés (*open source intelligence*). Las fuentes de información OSINT hacen referencia a cualquier información desclasificada y públicamente accesible en Internet de forma gratuita. Podemos afirmar entonces que cualquier Blog, página web de empresa, periódicos online, redes sociales, foros, e incluso bases de datos gratuitas, constituyen el grueso de estas fuentes de información.

Otros autores hacen evolucionar el concepto de inteligencia competitiva hablando de "inteligencia social" para referirse a la capacidad de recopilar, analizar y compartir conocimientos para mejorar la toma de decisiones, basadas en las interacciones en los medios sociales.

Gracias a su extraordinario potencial, las redes sociales presentan enormes posibilidades para todo: para el individuo como ciudadano del mundo, para el ciudadano como consumidor, para el individuo que tiene necesidades de comunicación, y para las empresas y organizaciones.

En un uso empresarial como herramientas de comunicación, la utilización estratégica de redes sociales permite:

- Proyectar la imagen de marca de la empresa: acercar la empresa al público objetivo (audiencia) y darse a conocer de forma informal y cercana, intercambiando, creando y compartiendo contenido y, al mismo tiempo, hacer ofertas y promociones sobre nuestros productos.

- Conversar de forma activa con los interlocutores: preguntar, escuchar y responder a clientes, proveedores, colaboradores, etc. y conversar

directamente con las personas que forman parte de la comunidad, principalmente con los clientes.

- Identificar las necesidades de nuestros potenciales clientes, conocer el grado de satisfacción sobre nuestros productos o los de nuestros competidores, establecer mejoras y realizar propuestas concretas.

- Establecer una red de colaboradores o *networking*: participando en redes especializadas en su ámbito sectorial o profesional, las empresas pueden conocer mejor a la competencia, compartir ideas e incluso entablar relaciones para desarrollar futuros proyectos conjuntos con otras entidades.

En cuanto al uso de las redes sociales en el proceso de búsqueda inteligente de información, un adecuado análisis y estudio de las redes sociales nos permitirá:

- Estar alerta de reacciones, comentarios y opiniones sobre nuestras marcas, productos y servicios y los de la competencia que nos permitirá actuar a tiempo ante incidencias (aspectos internos) y cambios en el entorno (aspectos externos).
- Identificar nichos de mercado donde podemos innovar, mejorando o lanzando nuevos servicios, productos o procesos.
- Detectar posibles focos de crisis que afecten a la empresa: incidencias en la atención al cliente, fallos de calidad en nuestros productos, etc.
- Conocer las acciones de marketing online de nuestros competidores y ver qué "valores" consiguen introducir respecto a su imagen y marca.
- Crear nuestras propias estrategias de marketing basadas en una segmentación adecuada de clientes. Estudio de perfiles de usuarios.
- Predecir y observar tendencias de consumo a partir de la opinión de consumidores y de las acciones y contenidos de marcas líderes (en nuestro sector o no).
- Buscar colaboradores para el desarrollo de proyectos conjuntos
- Descubrir nuevos procesos y productos, para aplicar esas innovaciones en nuestra organización.
- Conocer las tendencias del mercado y los cambios en otros sectores que pueden afectarnos.
- Conocer las demandas y nuevas necesidades de los consumidores y adaptar nuestra oferta como solución.
- Seleccionar y reclutar recursos humanos y hacer un seguimiento de la actividad profesional de posibles candidatos.

El estudio y análisis de las redes sociales debe hacerse desde una doble perspectiva:

- Por un lado, entendiendo las redes sociales como fuente de información. Ya no sólo las bases de datos, empresas, instituciones, etc. tienen posibilidad

real de publicar información y difundirla. Ahora todos los profesionales, expertos, consumidores, etc. publican y difunden. Cualquier acontecimiento, dato publicado, etc. se difunde con inmediatez, y por tanto las empresas pueden acceder a él a la hora de tomar decisiones. Por tanto, hace falta una capacidad de reacción más ágil para lograr anticiparse al competidor.

- Por otro lado, considerando el "efecto multiplicador" que poseen las redes sociales, como medio de difusión y que tiene que ver con el alcance y el impacto de la información. Esto puede producir diferentes efectos como son los siguientes:

    - Efecto eco: se trata del efecto que simplemente produce la repetición hasta el infinito de una información. Este efecto causa por un lado que la información sea conocida por otros, pero por otro lado conlleva el riesgo de perdurar en el tiempo más de lo necesario.
    - Efecto disonancia: se produce cuando se reciben muchos impactos informativos al mismo tiempo de forma que se produce una tensión en el receptor y un posible rechazo.
    - Efecto desconfianza: ya no todos los emisores son fiables. Algunos simplemente repiten lo que oyen sin comprobar si es verdad o no lo que dicen, otros no han profundizado lo suficientemente en la información y muchísimas no son fuentes primarias. Esto produce que el receptor muchas veces rechace al medio confundiéndolo con las fuentes que utilizan estos medios.
    - Efecto manipulación: las redes sociales son un medio perfecto para propagar informaciones falsas, sesgadas o manipuladas.

Siendo así, el análisis de redes sociales puede hacerse con diversas técnicas y herramientas según el objeto de análisis y según el tipo de información que deseemos obtener. Este trabajo se centra en el análisis de los textos de las conversaciones y para ello utiliza técnicas de **minería de textos.**

La minería textual es una aplicación de la lingüística computacional y del procesamiento de textos que pretende facilitar la identificación y extracción de nuevo conocimiento a partir de colecciones de documentos o corpus textuales. (Eíto Brun y Senso, 2004).

Las funciones que principalmente debería satisfacer una herramienta de minería textual, o el output que podemos esperar de ellas, incluiría:

- Identificar "hechos" y datos puntuales a partir del texto de los documentos.
- Agrupar documentos similares (*clustering*).

- Determinar el tema o temas tratados en los documentos mediante la categorización automática de los textos.
- Identificar los conceptos tratados en los documentos y crear redes de conceptos.
- Facilitar el acceso a la información repartida entre los documentos de la colección, mediante la elaboración automática de resúmenes, y la visualización de las relaciones entre los conceptos tratados en la colección.
- Visualización y navegación de colecciones de texto.

Las técnicas de Procesamiento del Lenguaje Natural (PLN), desarrolladas en el campo de la lingüística computacional están siendo ampliamente utilizadas en áreas tales como la minería de textos, de la recuperación de la información, *clustering* documental, análisis de opiniones o gestión del conocimiento.

El objetivo general de los sistemas de Procesamiento del Lenguaje Natural (PLN) es el tratamiento automático de la lengua a fin de ser interpretada y/o producida a la manera en la que lo hacemos los seres humanos. Se distinguen distintos aspectos interrelacionados:

- Fonética y fonología: el estudio del análisis de las señales acústicas que emitimos y generamos.
- Morfología: El estudio de la información contenida en una palabra considerada ésta en el contexto en el que se utiliza.
- Sintaxis: Estudio de las relaciones estructurales entre las palabras en una frase. Estudio de cómo ordenar y agrupar las palabras en la frase.
- Semántica: Estudio de los significados de las palabras y su forma de combinarse para formar significados más complejos.
- Pragmática, Discurso: Estudia cómo el contexto afecta a la interpretación de las oraciones.

En este trabajo nos vamos a centrar en el análisis semántico. El análisis semántico es el análisis del contenido que expresamos cuando escribimos un comentario. Varios son los análisis que podemos hacer: (Valderrábanos, 2001)

- Detectar nombres de cosas para seguir la marca o el producto que nos interesa. La semántica puede ser de gran ayuda para filtrar los documentos no relevantes, por su contexto o por su campo semántico.

- Además de nombres de cosas, podemos seguir temas, como “limitación de velocidad a 110”, “precio de la gasolina”, “ley Sinde”. Podemos vigilar la marca “Cepsa” y los temas que pueden estar relacionados, por ejemplo “emisiones de CO2”. Seguir temas es una tarea compleja. La semántica nos ayuda a seguir este tipo de temas, a pesar de que no se expresen de una sola manera (como “Mapfre”), sino de muchas, como “violencia de género”, “maltrato del cónyuge”, “atacó con un arma a su pareja”.

- Análisis de sentimiento: detectar el sentimiento, la actitud del que escribe el comentario, si es positiva o negativa; también conocido como “análisis de opinión”. Sin duda es útil saber si alguien habla del Peugeot, pero es mucho más útil distinguir si dice “me encanta el nuevo Peugeot” o “el nuevo Peugeot es lo peor”. Dando un paso más, existen muchas otras expresiones que tienen tanto o más interés para nuestro negocio y que no son opiniones, sino descripciones de hechos, como “me han perdido las maletas otra vez” o “nadie coge el teléfono en atención al cliente”. A pesar de que estas expresiones no sean opiniones (no son subjetivas, describen hechos), sí que reflejan lo que nuestro usuario o cliente piensa y por ello debemos capturarlo también; y a detectar estas construcciones también nos ayuda la semántica, yendo más allá de los simples diccionarios de valoración (“encantar”, positivo; “odiar”, negativo) y detectando construcciones complejas como “perder (verbo) + “maleta” (objeto directo). Además, y sin salir del terreno del sentimiento, es muy diferente la forma en que se expresa la crítica, por ejemplo, en una red social como Twitter y en un medio de prensa tradicional como Expansión o 5 Días. El primero es directo, “nunca había estado más harto de mi ADSL”, mientras que el segundo usa un tono mucho más neutro, más pegado a los hechos, “el número de reclamaciones ante consumo ha crecido en el sector de la telefonía”. Esto significa que tenemos que usar diferentes técnicas de análisis según el tipo de medio al que nos enfrentemos. Afortunadamente, la semántica también nos provee de soluciones para este tipo de necesidades.

- Categorizar los comentarios de los usuarios según áreas de negocio. El impacto que tiene un comentario negativo sobre un anuncio, por su música por ejemplo, es muy distinto del impacto de un comentario negativo sobre, por ejemplo, la calidad de un producto, o sobre el servicio de atención al cliente de la misma empresa. En primer lugar, afecta de manera muy diferente a la cuenta de resultados; en segundo lugar, debe llegar a diferentes departamentos de una misma organización para poder generar la respuesta que nos pide el cliente. La semántica nos ayuda a automatizar esta categorización de manera fiable y, sobre todo, flexible, adaptándose a la forma de organizar el negocio de cada empresa u organismo público, y a las propias características del producto o servicio. En otras palabras, no podemos trabajar con un conjunto de categorías únicas o cerradas, las categorías deben ser abiertas y adaptarse a las que cada entorno requiera.

- Descubrir información: como hemos visto, la semántica nos ayuda a encontrar lo que buscamos; además, puede ayudarnos a descubrir temas que no buscábamos y que pueden ser por eso los más relevantes para nosotros. Por ejemplo, saber que nuestros usuarios nos asocian con competidores de gama inferior a la nuestra, como empresas *low cost*, puede ser síntoma de un problema de posicionamiento. Dado que es algo que no esperamos, es algo sobre lo que no encargaremos nunca un estudio a nuestro departamento de marketing, ni siquiera se nos habrá ocurrido que tal cosa pudiera ocurrir. Por eso, es importante poder descubrir la información relevante, no meramente

encontrarla cuando la buscamos. Para ello, basta con aplicar una perspectiva *bottom-up*, descender al último dato (cada tweet, cada post…), identificarlo, capturarlo y analizarlo semánticamente, aislando cada átomo de contenido y detectando los rasgos semánticos compartidos por diferentes átomos para extraer puntos comunes y tendencias ("nueva pérdida de maleta", "perdieron mi equipaje otra vez", "otra vez sin maletas después de 18 horas de vuelo"…). En lugar de aplicar el enfoque *top-down* (con el que encontramos cosas que ya sabíamos que nos interesaban), podemos saber, por ejemplo, si se nos percibe como un producto caro, si hay quejas de las actitudes públicas de nuestros directivos, etc., sin tener que lanzar consultas sobre cada una de estas problemáticas. En otras palabras, puede ser muy productivo dejar de hacer preguntas dirigidas y simplemente escuchar para entender qué es lo que de manera espontánea nos están diciendo las redes sociales.

Las medidas basadas en coocurrencias consisten en un enfoque no supervisado para el cálculo de similitud semántica entre términos. Los valores de similitud son computados a partir del conteo de coocurrencias de términos en un corpus de documentos. En este sentido, constituyen una alternativa satisfactoria ante la carencia de recursos lingüísticos construidos por expertos. Por otro lado, la naturaleza de las relaciones representadas en las medidas de similitud puede ser ambigua. Términos con un valor alto de similitud pueden ser sinónimos, antónimos, o simplemente conceptos que aparecen a menudo en el mismo contexto. Es esa relación la que queremos descubrir.

Las técnicas de minería de textos y PLN se aplican en la monitorización de las redes sociales. En esta labor de automatización, hay dos puntos críticos:

- Cobertura: es importante poder analizar todo lo que nos interesa (lo que se dice sobre nosotros, sobre la competencia, sobre nuestro sector, sobre mercados en los que queremos entrar…). Interesa recoger y analizar todo, no sólo aquello que nos da tiempo a leer.
- Velocidad: es necesario hacerlo a tiempo, en tiempo real o casi; esperar semanas o meses puede quitarle sentido al análisis y a las decisiones que se pueden tomar a raíz de él.

Dado el ingente volumen de información textual generado a diario, se hace cada vez más necesario que el procesamiento y análisis lingüístico de esta información se efectúe de manera eficiente y escalable, lo que provoca que las tareas de PLN requieran de soluciones de tecnología *Big Data*.

La filosofía de los enfoques más recientes de la lingüística de corpus se basa en la *Web As Corpus*, línea de investigación donde se postula que con más datos y más texto se obtienen mejores resultados. Y para procesar más corpus a la escala de la Web se requieren soluciones HPC, como la elaborada por Gamallo et al., 2014.

A diferencia de los algoritmos de minería de datos donde existen herramientas específicas que explotan las capacidades analíticas de Hadoop (p.ej. Apache Mahout para clasificadores, recomendadores y algoritmos de *clustering* y Apache Giraph para el procesamiento de grafos), no conocemos a día de hoy ninguna herramienta que emplee de forma integrada soluciones de PLN en Big Data.

Recientemente, el paradigma MapReduce se ha comenzado a aplicar a algunas tareas de PLN, como por ejemplo la traducción estadística (Ahmad et al., 2011; Dyer, Cordora, y Lin, 2008), la construcción de matrices de coocurrencias (Lin, 2008), la minería de textos (Balkir, Foster, y Rzhetsky, 2011), la computación de similitudes semánticas (Pantel et al., 2009), o la adquisición de paráfrasis (Metzel y Hovy, 2011) (en Gamallo et al., 2014).

La principal aportación de este trabajo es la incorporación de la variable "tiempo" (fecha de publicación del texto) a los dos tipos de análisis planteados, el análisis de palabras coocurrentes y la cohesión léxica. Esto permitirá hacer análisis de la evolución temporal de las relaciones entre los términos y también permitirá observar desde una perspectiva temporal el grado de cohesión que mantienen los mensajes en diferentes periodos.

Para la introducción de la variable temporal será necesaria la creación de matrices de coocurrencia para cada periodo de tiempo analizado. No se trabajará con una matriz de coocurrencia que recoja la totalidad de documentos del corpus de trabajo, sino que se elaborarán matrices para cada grupo de documentos incluidos en el marco temporal que se haya determinado para hacer el análisis (puede ser un día, un mes, o un periodo específico). La construcción de matrices de coocurrencia tiene un coste operacional y de computación muy bajo. La aplicación de técnicas *MapReduce* mencionadas anteriormente (Lin, 2008) podría favorecer la aplicación de este método para el procesamiento de información en tiempo real para poder hacer análisis en marcos temporales más reducidos (por horas, por ejemplo). Veremos en este trabajo las aplicaciones prácticas de este tipo de análisis.

2. Estado del arte

El análisis de la coocurrencia de palabras estudia la aparición conjunta de dos o más palabras en unidades textuales. Si la coocurrencia entre dos palabras es elevada, significará que existe una importante proximidad o relación entre ambas palabras. Esta proximidad puede cuantificarse mediante diversos índices y métricas, y dibujarse obteniendo así grafos de palabras o redes de palabras. Cuanto más juntas están dos palabras en el mapa, mayor es la relación entre ellas.

Sin duda, una de las aplicaciones más extendidas del análisis de coocurrencias es la detección de temas mediante la creación de matrices de coocurrencia y aplicación de algoritmos de *clustering* o agrupación. En eso se basa el "método de palabras asociadas" que se trata más adelante con más detalle puesto que es la base de la propuesta de análisis de este trabajo.

El análisis de palabras asociadas, o método de palabras asociadas (*coword analysis*, en inglés) es una metodología ciencimétrica que permite dibujar redes de palabras a partir de documentos textuales. Se emplea en los sistemas de conocimiento como instrumento para identificar los centros de interés o temas de un campo científico. Fue ideado a principios de la década de los 80 en el "Centre de Sociologie de l'Innovation de L'École de Mines" de Paris por Michel Callon, John Law y Jean Pierre Courtial.

La red o mapa generado se divide en temas de investigación, definidos por sus descriptores y los parámetros de densidad y centralidad. Igualmente permite disponer estos temas en un diagrama estratégico que nos representa tanto el desarrollo interno de los temas como su capacidad de relación con los demás dentro de la red científica. (Ruiz-Baños y Bailón Moreno, 1998)

Para la puesta en marcha de este método se desarrolló un conjunto de programas informáticos denominado LEXIMAPPE. Leximappe[1] se aplica a todo tipo de documentos indizados mediante palabras clave y en especial a los artículos científicos y técnicos, patentes… De forma más general es a aplicable a cualquier documento textual, siempre y cuando se haga una indización semiautomática previa mediante un programa adecuado.

En 1997 los autores Van Meter y Tumer (1997) basaron su trabajo en el método de palabras asociadas. El principal objetivo de era detectar, a partir de los conceptos principales utilizados dentro de una determinada disciplina académica, diferentes grupos de investigadores, así como la propia evolución de la disciplina. Concretamente, estos autores analizaron las coocurrencias de palabras clave en los resúmenes cortos aparecidos en las reseñas de "*Sociological Abstracts*" relacionadas con los temas del SIDA (Van Meter y Tumer 1997) y el capital social (Van Meter 1999). Utilizando combinadamente los programas Lexinet y Leximappe se identifican las coocurrencias de palabras clave, a continuación se calcula un índice de proximidad,

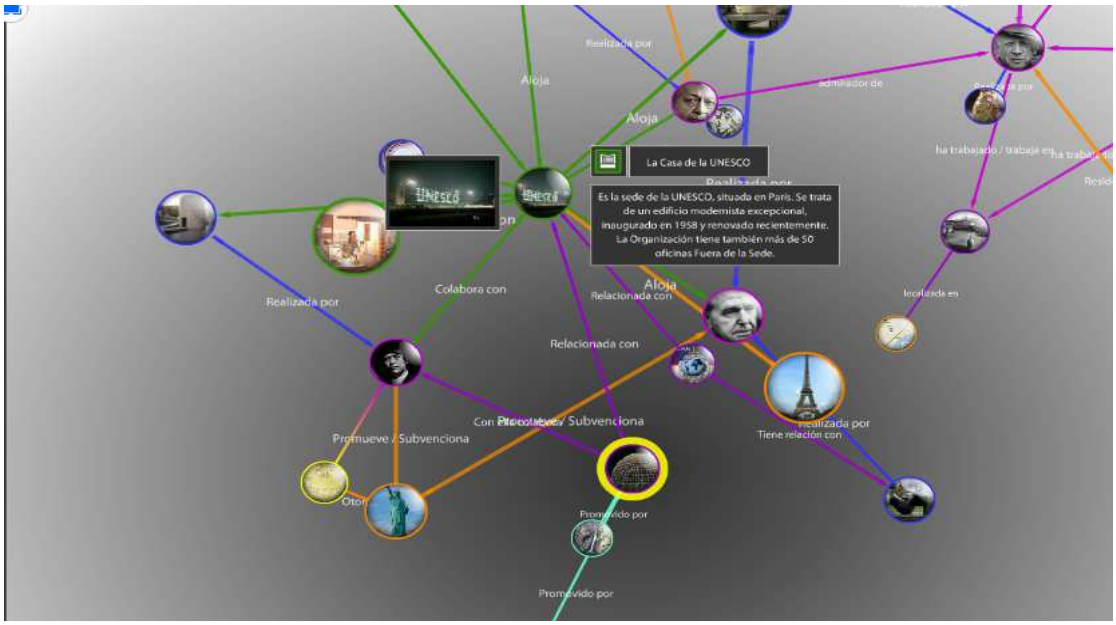
[1] No se ha encontrado sitio web oficial de Leximappe y la última referencia encontrada sobre el software es de 2010

basado, obviamente, en el número de coocurrencias entre ellas, y a continuación se utiliza este índice para construir conglomerados de palabras con un máximo de diez por grupo. Estos grafos se representan posteriormente sobre dos ejes, siendo el primero un indicador de centralidad (la media de los vínculos entre el conglomerado y las palabras fuera de él), y el segundo un indicador de densidad o cohesión interna del grupo (la media de la intensidad de asociación entre las unidades del conglomerado). La construcción de grupos y el cálculo de los indicadores de centralidad y densidad de cada uno de ellos permite a Van Meter y Tumer describir la evolución de los contenidos de la actividad científica en que están interesados, así como las interrelaciones existentes entre los diferentes temas.

Como afirman Ruiz-Baños y Bailón Moreno (1999), el método de palabras asociadas es capaz no sólo de descubrir la estructura interna de las redes científicas sino también de poner de manifiesto la dinámica evolutiva de esta estructura, en la línea en la que lo utilizan Van Meter y Tumer. Los temas de investigación se van encadenando de generación en generación formando las denominadas series temáticas. Estas series temáticas son representables sobre un diagrama cronológico, un diagrama estratégico y seguirse en la evolución de sus parámetros o ciclos de vida. En el apartado correspondiente a la explicación del método veremos las herramientas que utiliza para abordar adecuadamente este aspecto.

El análisis de coocurrencias se ha utilizado especialmente en el tratamiento y la gestión de grandes colecciones de textos, principalmente del ámbito científico. El principal objetivo de la aplicación de las técnicas de minería de textos, y concretamente el análisis de coocurrencias en estos casos es la representación de datos en **mapas de conocimiento**. Dado el montante de información científica que se produce, las técnicas tradicionales de extracción de conocimiento, como la lectura, están dejando paso a otras formas de análisis. Mediante técnicas de minería de textos sobre datos bibliográficos se puede mostrar en los mapas de conocimiento el estado de un área de conocimiento, se pueden generan indicadores científico técnicos o se experimenta con técnicas para conocer las necesidades de los usuarios.

Los mapas de conocimiento son instrumentos que ayudan a visualizar información y resultan de gran ayuda para identificar y representar los recursos de conocimiento de los que dispone una organización. El *text mining* puede realizar cálculos matemáticos y estadísticos sobre algunos campos de los artículos para revelar sus patrones o estructura interna, así como la evolución de una determinada disciplina. Esta metodología aumenta la objetividad de las conclusiones que se derivan de la interpretación de los mapas. (Garcia, A. et al., 2015)



Ejemplo 1: Mapa del conocimiento de las obras de arte y proyectos especiales de la Unesco http://www.millansoft.es/gestion-del-conocimiento/representacion/ejemplo-obras-de-arte-de-la-unesco.html

En la misma línea, se habla de **mapas tecnológicos**, que se realizan también mediante el análisis de coocurrencias. Los mapas tecnológicos son representaciones visuales del estado de la tecnología en un ámbito o área determinados. Los mapas presentan gráficamente, de forma sintética, las tecnologías en que se ha investigado más y, en consecuencia, publicado y patentado más en un período determinado. Permiten también detectar aquellas tecnologías emergentes que están experimentando una rápida expansión mediante la comparación con mapas correspondientes a períodos anteriores.

Por ejemplo, en una base de datos de artículos sobre superconductividad, el análisis de coocurrencia pretende detectar cuántas veces las palabras "bario" e "itrio" aparecen juntas en los títulos. Si la coocurrencia es elevada, es decir si el número de veces que "bario" e "itrio" figuran juntos es alto respecto al número total de artículos considerados, significará que existe una importante proximidad o relación entre ambas palabras. Por el contrario, una coocurrencia baja o nula entre dos palabras será señal de una falta de relación o lejanía entre ellas. Esta proximidad o lejanía puede cuantificarse mediante diversos índices y métricas, y dibujarse obteniendo así los mapas tecnológicos (Callon, Courtial y Penan, 1993, Escorsa, Maspons y Rodríguez, 1998, Escorsa y Maspons, 2001 EN Escorsa).

Existen otras posibilidades de análisis de coocurrencias entre indicadores que pueden ser o no de la misma naturaleza: coocurrencias entre palabras clave de productos y/o tecnologías y empresas (que permite detectar en que productos y/o tecnologías trabajan las empresas de un sector), empresas-clases de la Clasificación Internacional de Patentes (para conocer las áreas en que está patentando cada empresa), productos/tecnologías-grupos de patentes, palabras clave-países, etc.

La **recuperación de la información** es otro de los campos en los que este análisis tiene muchas aplicaciones. Por ejemplo, en los programas de indización automática, en los que la indización es llevada a cabo por algoritmos que mediante diversas técnicas o métodos determinan cuál es el peso con el que cada uno de los términos que aparecen en el documento representa su contenido temático. En la llamada "Indización de Semántica Latente", en el marco de las **búsquedas semánticas**, un documento puede ser descrito por términos que no aparecen en el documento, pero que presentan fuertes relaciones de coocurrencia con los términos del documento en otros documentos de la colección.

Los grafos de coocurrencia léxica también han sido utilizados en lingüística computacional en experimentos de **desambiguación de sentidos** (Fernandez-Amoros et al., 2010) y para la creación de **taxonomías** mediante el tratamiento de términos hiperónimos (Nazar, R. et al, 2012)

Una de las formas de análisis y representación de las coocurrencias son las llamadas "redes de palabras".

Una red de palabras es la transformación de un texto, escrito en un determinado lenguaje, en un grafo G(N,E), donde G es el grafo, o red, compuesto por N palabras (o términos) distintas y E enlaces que las relacionan en un texto (Cárdenas, Losada, Moreira, Torre y Benito, 2011).

En este trabajo se utilizarán redes de palabras para representar el contenido de los textos analizados, basándose en la copresencia de los términos que aparecen con más frecuencia. Verd Pericas (2005) explica:

En la construcción de redes de palabras pueden observarse dos procedimientos principales.

- *En primer lugar, aquellas que se construyen sobre la base de una matriz de actores por palabras (matriz de afiliación), de la que puede obtenerse posteriormente una matriz valorada de palabras por palabras (matriz de adyacencia) que representa el número de ocasiones en que cada una de ellas coocurre en un mismo actor.*
- *En segundo lugar, otro modo de construcción de redes de palabras es la utilización de algún método de codificación automatizada que tenga en cuenta las coocurrencias de unas mismas palabras o conceptos en un espacio textual de una amplitud dada dentro del texto o conjunto de textos analizados. En este caso la relación que une las palabras es la copresencia en la «ventana» de texto considerada.*

I

La principal ventaja del trabajo con redes de palabras es la simplicidad de los procedimientos y la capacidad insuperable de analizar grandes cantidades de texto, motivo por el cual este trabajo se aborda desde esta perspectiva.

Verd Pericas (2005) hace un repaso de diversas metodologías en el tratamiento de redes de palabras que se basan en la copresencia de términos.

Los trabajos de Danowski (1988, 1993) son los que mejor ilustran los procedimientos de análisis basado en redes de palabras mediante el uso de la codificación automatizada. "*Word-Network analysis*" es el término con el que denominó su procedimiento y que se ha llegado a extender para referirse a todo el conjunto de aproximaciones similares.

El procedimiento (Danowski 1993:203-15) consiste en crear una matriz de palabras por palabras a partir de la coocurrencia en una «ventana corredera» (o fragmento del texto estudiado) de entre 7 y 11 palabras de amplitud.

El origen de los datos puede ser muy variado: mensajes electrónicos entre miembros de una misma empresa, noticias aparecidas en los periódicos, transcripciones de entrevistas y grupos de discusión, argumentos de libros...

La red de palabras que se obtiene como resultado puede analizarse de diferentes modos. Mientras Danowski se orienta principalmente hacia un análisis estructural, en el sentido de que normalmente presta más atención a los grupos de palabras que se forman y a las distancias existentes entre palabras que a la posición concreta de ciertas palabras en relación a las otras (1988:412-17); Schnegg (1997) estudió las diferencias culturales entre 35 empresas norteamericanas y japonesas a partir del análisis de las palabras más utilizadas en una selección de informes públicos y su análisis se centra especialmente en la «comunidad semántica» de las palabras más utilizadas en cada grupo de empresas, investigando también a través de qué palabras se produce la relación entre las diferentes comunidades semánticas (es decir, entre los términos más utilizados por cada grupo de empresas).

Los trabajos realizados por el equipo formado en tomo a Gorman (Brandes y Gorman 2003, Gorman y Dooley 2001, Gorman et al. 2002) constituyen un distanciamiento respecto al modelo de «ventanas correderas». Realizan un análisis denominado *Centering Resonance Analysis* (CRA) basado en la teoría lingüística de los «centros conversacionales»; el objetivo principal de este tipo de análisis es el de localizar las palabras con una posición estructural más relevante en un conjunto dado de textos Los nombres y adjetivos presentes en los sintagmas nominales constituyen los nodos de la red, siendo la relación que los va a unir el hecho de formar parte del mismo sintagma nominal o de sintagmas nominales consecutivos en la misma fijase. La red de palabras resultante contiene diferentes intensidades en las relaciones, en función del número de veces que coocurren las palabras. Esta aproximación se ha aplicado a diferentes tipos de textos: análisis de las informaciones que tras los atentados del 11 de septiembre de 2001 estuvo difundiendo la agencia de noticias Reuters (Gorman y Dooley 2(X)1), análisis de transcripciones de interacciones lingüísticas entre dos grupos enfrentados (Gorman et al. 2(X)2), y evaluación de las percepciones que un grupo de estudiantes tenían de dos textos diferentes (Gorman et al. 2002).

Finalmente, como muestra del procedimiento que construye las relaciones entre palabras basándose en la copresencia en un mismo emisor mencionamos aquí el

trabajo de Schnegg y Bemard (1996). El citado trabajo tenía como objetivo conocer qué conceptos representaban mejor las motivaciones centrales de un conjunto de estudiantes de antropología para elegir dicha titulación. De las transcripciones de las entrevistas se obtuvo un listado de palabras depurado, representativas de los temas; se realizó una matriz de coocurrencia a partir de los valores de la cual se fueron aplicando valore de corte para finalmente obtener diferentes grupos de palabras representativas de las motivaciones de los estudiantes. Los resultados muestran un núcleo central de palabras común a todos los entrevistados, lo cual, según los autores, es muestra del elevado consenso existente entre los informantes.

En la línea planteada por Schnegg y Bemard (1996) (en Verd Pericas, 2005), una de las posibilidades del análisis de redes de palabras basadas en la copresencia de los términos, es la de observar el **grado de cohesión** de los diferentes mensajes para un grupo de personas o para un grupo de textos.

Según Melgar (2011), "en relación a la definición de la cohesión textual, se confrontan dos concepciones fundamentales: La de aquellos que perciben la cohesión como el conjunto de relaciones semánticas que permiten interpretar el significado de un elemento en el texto a partir de otro elemento que aparece anterior o posteriormente en el contexto lingüístico del texto, una postura que hace innecesaria la definición de coherencia al establecer la naturaleza semántica de dichas relaciones, caso de Halliday & Hasan (1976). Y la de aquellos que entienden la cohesión como el conjunto exclusivo de recursos formales a nivel de superficie del texto, que reflejan la coherencia de la estructura profunda, caso de Beaugrande & Dressler (1981)"

Según Halliday y Hassan (1976), hay cinco tipos de enlaces cohesivos: sustitución, elipsis, referencia, conjunción y cohesión léxica:

> *La cohesión léxica consiste en conectar semánticamente expresiones léxicas relacionadas, bien sea por identidad, oposición, inclusión o con continuidad, con el fin de referirse a una misma realidad sin emplear necesariamente las mismas palabras. Incluye dos procedimientos principales: la reiteración y la coocurrencia. La reiteración, a su vez, agrupa la sinonimia, la repetición, la superordenación y la generalización. Por su parte la coocurrencia tiene que ver con la forma como las palabras se vinculan unas con otros, estableciendo redes de significado que le dan sentido al texto.*

Como vemos, aparece el análisis de coocurrencias como uno de los procedimientos para medir la cohesión léxica de los textos.

Las definiciones de cohesión textual y léxica se restringen a los textos individuales pero en este trabajo se va a aplicar a la comparación de diferentes textos. Se pasa del nivel intratextual al intertextual.

Así explican el concepto de "intertextualidad" Keith y Trellis (2006):

> *Autores como Kristeva (1966), Barthes (1970) o Bajtín (1986) entienden la intertextualidad en el sentido de que un texto siempre está vinculado a textos o experiencias previas y muestran prospección a textos o enunciados futuros. De Beaugrande y Dressler (1981) afirman que cualquier texto debe cumplir con el requisito de la intertextualidad para que pueda considerarse como texto, que, además, determina la manera en que el uso de un cierto texto depende del conocimiento de otros textos. Para estos autores, el término intertextualidad se refiere a la relación de dependencia que se establece entre, por un lado, los procesos de producción y recepción de un texto determinado y, por otro, el conocimiento que tengan los participantes en la interacción comunicativa de otros textos anteriores relacionados con él. Según Stubbs (1996), los textos se orientan conforme a rutinas y convenciones; están modelados por textos previos a los que hacen referencias intertextuales, probablemente incluidas en el mismo corpus. En este sentido, Stubbs (2001: 120) señala, "Analysis cannot be restricted to isolated texts. It requires an analysis of intertextual relations, and therefore comparison of individual instances in a given text, typical occurrences in other texts from the same text-type, and norms of usage in the language in general". En esta misma línea, encontramos la postura de Fairclough (2002), quien defiende una perspectiva intertextual para el análisis, por ejemplo, de frases pre-construidas y colocaciones fijas.*

Keith y Trellis (2006) hablan de "densidad conceptual" para referirse al análisis de las redes de palabras teniendo en cuenta sólo aquellas combinaciones que son más fuertes, es decir, más frecuentes, y por tanto supone que tienen entre sí una relación más estrecha o significativa. El análisis opuesto sería el de visualizar todas las relaciones que un término presenta, obteniendo la "dispersión léxica" que indica mayor complejidad de la red pero con vínculos menos consistentes.

Marín (2013) introduce una nueva interpretación. El autor habla de cohesión y temperatura del lenguaje.

- La métrica de la cohesión (que corresponde con la densidad, es decir, la relación de los términos dentro del grupo), pretende medir la "unicidad" o cohesión del mensaje: mensajes semánticamente muy parecidos utilizan conceptos muy similares mientras que los mensajes semánticamente diferentes utilizan conceptos dispares. Se trata de ver los conceptos del mensaje como una red cuyos enlaces son proporcionales al número de veces que coocurren dichos conceptos en un mismo mensaje. Un conjunto de mensajes cuyo contenido sea exactamente el mismo, daría como resultado una cohesión equivalente a la unidad.

- La temperatura del vocabulario hace referencia al ritmo con el que se crea un lenguaje. Así, en un periodo donde circulan mensaje con conceptos muy

> novedosos respecto al periodo anterior se obtienen temperaturas léxicas muy altas. Por otro lado, un periodo donde los mensajes son exactamente iguales a los aparecidos en el anterior, tendría una temperatura exactamente igual a cero.

Estos dos conceptos nuevos que introduce Oscar Marín (2013) suponen la base de la propuesta de este trabajo porque introduce la variable temporal en el análisis de coocurrencias y de cohesión léxica. Sobre esta base, se aplica el método de palabras asociadas que aporta la metodología estadística necesaria para el trabajo con palabras coocurrentes y su evolución. Este método se ha aplicado en numerosas investigaciones científicas pero no ha sido adaptado al procesamiento del lenguaje natural y no se ha aplicado a textos cortos como son los de la red social Twitter.

3. Propuesta metodológica: incorporación de la variable temporal al análisis de coocurrencias y cohesión léxica

3.1 Objetivo

En este trabajo se expone una aproximación al análisis de coocurrencias y cohesión léxica cuya principal aportación es la introducción de la variable temporal. Para este objetivo se aplica como base el método de palabras asociadas (*Co-word method*) aunque se introducen ligeras modificaciones para adaptarlo a nuestro objetivo.

3.2 Método de palabras asociadas

Para el análisis se aplica una adaptación del método de palabras asociadas (MPA) descrito por Courtial en 1990.

Descripción del método:

3.2.1 Cálculo del coeficiente de asociación

A partir de una matriz de coocurrencia, se calcula, para cada par de palabras el índice de asociación E_{ij} dado por:

$$E_{ij} = \frac{c^2_{ij}}{c_i c_j}$$

Donde C_{ij} es el cuadrado de la coocurrencia entre la palabra i y la palabra y C_i y C_j son las frecuencias absolutas de cada una de las palabras en de la matriz de coocurrencias.

El coeficiente de asociación es el producto entre la probabilidad de que aparezca la palabra *j* cuando se presenta la palabra *i* y la probabilidad de tener la palabra *i* cuando se presenta la palabra *j,* por lo cual varía entre cero y uno. Es un índice de similitud entre las palabras clave (en nuestro caso, pares de palabras más frecuentes) que se puede utilizar para la aplicación de métodos de clasificación (Charum, 1998).

Este índice de similitud o coeficiente de asociación, al que llamaremos *E*, muestra que dos palabras clave se encuentran cercanas en la medida en que aparezcan simultáneamente en un gran número de documentos, así, en el MPA se realiza una clasificación jerárquica mediante enlace simple, por lo que dos palabras se agrupan si son las más cercanas en términos de su asociación.

Con el cálculo de este índice se obtiene una matriz (matriz *E*) que puede interpretarse como un grafo donde los nodos son las palabras y sus vínculos son las asociaciones

entre ellas. El método corta ese grafo en subconjuntos de palabras relacionadas entre sí.

3.2.2 Índices de densidad y centralidad

Siguiendo el método MPA, la caracterización interna de los grupos de palabras se hace a partir de las relaciones internas de cada grupo (densidad) y de las relaciones entre grupos (centralidad).

La **densidad** es una medida de la fuerza de las asociaciones internas de un grupo o cluster. Se define como el promedio de los coeficientes de asociación entre las palabras dentro del grupo. Así, si *S* es un grupo creado, entonces su densidad D_s es:

$$D_s \quad \frac{1}{m^1} \quad \sum_{i \in S} \quad \sum_{j \in S} \quad E_{ij}$$

Donde m1 es el número de coeficientes de asociación internos no nulos.

De esta forma, si las palabras dentro de un grupo aparecen con alta frecuencia de forma simultánea en diferentes documentos, significa que el grupo está representando a una temática elaborada y tendría una densidad alta. Por otro lado, si las palabras dentro del grupo están presentes de forma simultánea sólo en algunos documentos, pero además se encuentran en otros documentos asociadas con otras palabras, se dice que el grupo representa a una temática poco elaborada y por lo tanto, su densidad es baja.

La densidad es importante en el momento de caracterizar un grupo de palabras, porque refleja si la temática que evidencia el mismo, está desarrollada o no.

La **centralidad** mide el nivel de relación de un grupo con los demás. Se calcula como el valor medio de los coeficientes de asociación entre las palabras clave de un grupo con las palabras clave que pertenecen a los demás grupos existentes. Es decir, como su nombre lo indica, la centralidad muestra la importancia de la temática en general. Si *S* es un grupo creado, entonces su centralidad C_S es:

$$D_s \quad \frac{1}{m^{11}} \quad \sum_{i \in S} \quad \sum_{j \in S} \quad E_{ij}$$

Donde m^{11} es el número de coeficientes de asociación externos no nulos

Si un grupo tiene un índice de centralidad alto, significa que la temática representada por éste tiene un alto impacto sobre las demás temáticas, por otro lado, si sucede lo contrario, la temática es poco central.

La interpretación que en la exposición del método se hace de estos índices de densidad y centralidad es fácilmente adaptable al trabajo con pares de palabras.

3.2.3 Creación de un diagrama estratégico.

El mapa estratégico es un diagrama de dos ejes donde se representan los grados de densidad (eje vertical) y centralidad (eje horizontal). En el diagrama estratégico cada grupo queda reducido a un punto en el plano, el cual se puede nombrar con la palabra cuya suma de asociaciones internas sea más alta.

Figura 1: Representación del diagrama estratégico

3.2.4 Grafos de palabras.

El análisis interno de cada grupo se hace dibujando el grafo o red de las asociaciones entre las palabras clave que pertenecen a cada uno de los grupos. Cada vínculo representa la asociación entre cada par de palabras clave (términos más frecuentes para este trabajo), por lo tanto, aunque la red de relaciones muestre diferentes uniones, no necesariamente esas uniones son iguales o tienen la misma intensidad para cada par.

3.2.5 Evolución temporal.

El método de las palabras asociadas también es capaz de poner de manifiesto las transformaciones temporales de la estructura interna de los textos. Las herramientas del MPA para abordar adecuadamente la cuestión de la evolución cronológica. (Ruiz-Baños, 1998)

- **Comparación de los temas**:

  - Índice de intersección: Número de palabras comunes que hay entre ambos temas. Supongamos dos temas T1 y T2 y queremos determinar su similitud. Definiremos índice de intersección como el número de palabras comunes, W12, que hay entre ambos temas (Callon, M.; Courtial, J. P. y Laville, F., 1991). Normalmente diremos que dos temas están relacionados por su similitud temática si su índice de intersección supera un umbral mínimo de, por ejemplo, 3. Este índice no es suficientemente ecuánime, ya que dependiendo del tamaño de los temas que se comparan, el número de palabras comunes puede representar fracciones de tema muy distintas y por tanto similitudes relativas variables: dos temas de 4 palabras en total con 3 comunes son, por supuesto, más similares que dos temas de 15 palabras en total y también con 3 comunes.

  - Índice de transformación: tenemos los temas T1 y T2 y queremos determinar cuánto se diferencian entre ellos. Para ello podemos definir el índice de transformación, *t*, como el cociente entre la suma de palabras existentes en ambos temas y el número de palabras comunes:

$$t = \frac{W_1 + W_2}{W_{12}}$$

Donde:

W_1.- Número de palabras del tema 1.

W_2.- Número de palabras del tema 2.

W_{12}.- Número de palabras comunes entre los temas 1 y 2

Hay que hacer notar que si dos palabras aparecen en los dos temas a la vez, deben contarse dos veces (CALLON, M.; COURTIAL, J. P. y LAVILLE, F., 1991). Cuando dos temas son iguales el índice de transformación vale 0 y cuando son totalmente distintos, infinito.

- Índices de influencia y de procedencia: miden el grado de continuidad entre las generaciones de tema.

    - Índice de influencia: es la proporción de palabras de un tema que reaparecen en otro tema de la siguiente generación.
    - El índice de procedencia muestra la proporción de palabras de un tema de segunda generación que proceden de un rema de primera generación.

- **Series temáticas**: una serie temática es un conjunto de temas de generaciones encadenados por un valor de similitud umbral. Esta relación viene determinada por el índice de similitud dinámica (ISD). El ISD es una medida de la cantidad de significación que un tema conserva a lo largo de las traducciones (evoluciones) que sufre a lo largo del tiempo.

- **Evolución de los descriptores de los temas**: los descriptores cambian con el tiempo, desapareciendo unos para dar paso a otros.

- **Movimiento de los temas en el diagrama estratégico.** El seguimiento de las posiciones de los temas en el diagrama estratégico nos permite establecer pautas de comportamiento evolutivo.

- **Ciclo de vida de los temas**: Supongamos que seguimos la evolución de un tema durante un periodo largo de años a través del estudio de sus propiedades. Tendremos sobre la serie temática informaciones abundantes y muy significativas. Esta evolución dinámica de las propiedades de un tema es lo que denominamos de un tema es lo que denominamos ciclo de vida. En una representación gráfica podemos incluir variables como las siguientes:

    - Índice de transformación que expresará los cambios conceptuales del tema.
    - Centralidad y densidad: nos ofrecerá una visión cuantificada y progresiva de la cercanía o alejamiento al centro de la red y del desarrollo interno.
    - Número de artículos: nos proporciona información sobre el tamaño que adquiere en cada momento el tema.

La aplicación de estas técnicas dependerá de los objetivos de análisis y del tipo de texto con el que se trabaje.

3.3 Adaptación del método de palabras asociadas. Incorporación de la variable tiempo

Como se ha explicado al principio de este punto, se trabaja con el método de palabras asociadas pero introduciendo algunas modificaciones necesarias para adaptarlo a los objetivos de este análisis.

Este método ha sido aplicado a diversos trabajos de carácter científico, pero no en textos escritos en lenguaje natural y de menor extensión, como son los mensajes de Twitter que se utilizan en este trabajo a modo de ejemplos. Sin embargo, consideramos que la adaptación a este tipo de análisis es posible mediante la substitución de palabras clave por pares de palabras coocurrentes, y mediante una interpretación diferente de los resultados, puesto que ya no se trabaja con grupos de temas.

El otro motivo fundamental por el que se decidió aplicar este sistema a nuestro estudio es que el método de palabras asociadas además define un conjunto de herramientas adecuadas para representar las transformaciones que sufren las redes de palabras a lo largo del tiempo.

La principal variación respecto al método original es que el MPA está ideado para la identificación de las diversas temáticas dentro de un conjunto de documentos a partir de las palabras clave que coocurren de forma simultánea. La extracción de temas no es el objetivo de este trabajo, por lo que no trabajaremos con palabras clave sino con los pares de palabras más frecuentes. Por lo tanto, no se aplican los algoritmos de *clustering* o agrupación automática.

La variable tiempo se introduce desde el principio. Los documentos se agrupan por fecha y se escoge el marco temporal en el que se quiera hacer el análisis (un mes, un día…). Se elabora una matriz de coocurrencia por día y se calcula el índice de asociación para cada par de palabras de todas las matrices.

En este punto, ya se puede hacer un primer análisis. Se puede calcular para cada par de palabras su posición en los diferentes marcos temporales, comprobando la velocidad con la que un par de palabras se mueve en el tiempo: velocidad positiva si ha pasado a ser más frecuente o velocidad negativa si es menos frecuente. Esto se puede representar en un gráfico de líneas, como el que sigue.

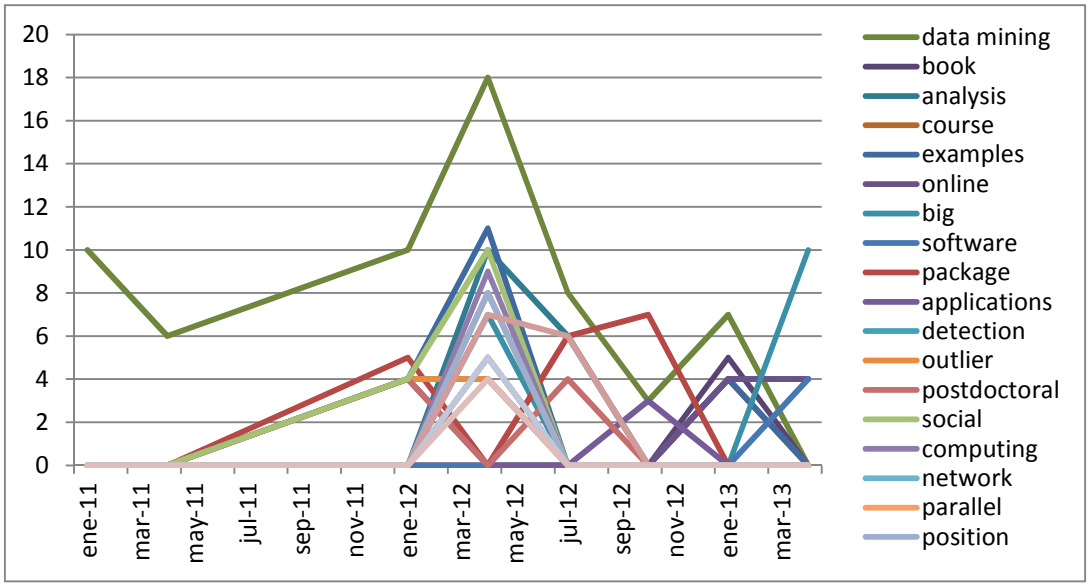


Gráfico 1: Las líneas muestran la frecuencia de aparición de los términos que coocurren con "R" (lenguaje de programación) en un periodo de 3 años. (Elaboración propia).

La representación temporal permite comparar la evolución de los términos de manera conjunta. En este caso, los tuits son un conjunto de documentos obtenidos de una búsqueda sobre el lenguaje de programación R. Puede hacerse sobre el conjunto de las palabras analizadas, como el caso presentado en el gráfico, o puede hacerse con pares de palabras individualmente. Se pueden analizar tendencias de crecimiento o decrecimiento, ver la evolución conjunta de determinados términos o al revés, ver si la el movimiento es opuesto y se debe a algún motivo.

Respecto a las relaciones entre los términos, el método de palabras asociadas contempla la creación de redes de palabras. Para el análisis de evolución se considera que estos gráficos no son necesarios, aunque serían un buen complemento para análisis más extensos.

Como hemos comentado, para la introducción de la variable tiempo, se agrupan los documentos por marcos temporales. Estos grupos son los equivalentes a los *clusters* del MPA.

El cálculo de los índices de densidad y centralidad no varía respecto al método original ya que se aplican de igual manera al conjunto de palabras de cada marco temporal.

El índice de **densidad** marca, según el MPA la fuerza de las asociaciones internas. Si este índice es muy alto, indica que la cohesión del conjunto de textos es alta, que hay una elevada sincronización. Si los conceptos expresados en los mensajes son muy dispares, el índice de densidad será muy bajo. Básicamente el índice de densidad para cada grupo indicará la fuerza de las relaciones entre los términos, lo que significa que coocurren con frecuencia, si el grado de densidad es alto y por tanto señalan que el mensaje se mantiene sincronizado.

Gráfico 2: Evolución de la cohesión léxica. Oscar Marín, 2014. Representación del grado de cohesión léxica de los mensajes de Twitter sobre el 15M en los días previos, centrales y posteriores al acontecimiento

Respecto al índice de **centralidad**, el MPA lo define como el nivel de relación de un grupo con los demás. Si el grado de centralidad es alto supone que la temática tratada tiene un alto impacto sobre los demás; por el contrario, si el grado de centralidad es bajo, la temática representada por ese grupo es poco central o significativa.

Desde la perspectiva de este trabajo, si el grado de centralidad del grupo de palabras (las correspondientes a un marco temporal) es muy alto, indica que los conceptos que trata no son nuevos, que están muy relacionados con los estudiados en el marco anterior, lo que quiere decir que mantienen la cohesión en ese sentido. Si por el contrario el índice es muy bajo, indicará que los términos utilizados o son bastante novedosos o bien están en declive.

Esto es lo que Oscar Marín (2014) denomina “temperatura léxica” y es la aportación que su trabajo hace al análisis de coocurrencias con el MPA. Así pues, en un período donde circulan mensajes con conceptos muy novedosos respecto al período anterior, se obtienen temperaturas léxicas muy altas. En cambio, en un periodo donde los mensajes son exactamente iguales a los aparecidos en el anterior, tendría una temperatura léxica igual a cero.

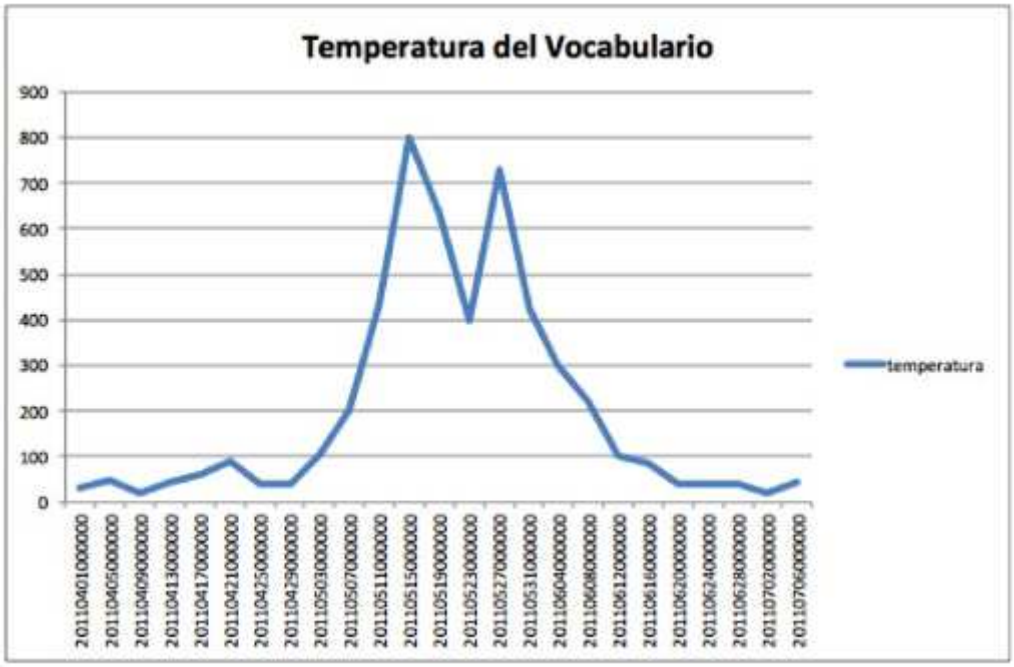


Gráfico 3: Evolución de la temperatura del vocabulario. Oscar Marín, 2014. Representación de la temperatura del vocabulario de los mensajes de Twitter sobre el 15M en los días previos, centrales y posteriores al acontecimiento

Para el análisis de la cohesión léxica se dibujan los índices en los mapas estratégicos del MPA donde ambos índices, centralidad y densidad, aparecen representados. Los grupos se sitúan entre los cuadrantes del eje bidimensional.

Figura 2: Diagrama estratégico.
Fuente:http://www.ugr.es/~rruizb/cognosfera/sala_de_estudio/ciencimetria_redes_conocimiento/estrategico.htm

La situación de los grupos de términos en uno u otro cuadrante, según el método de las palabras asociadas, tiene la interpretación apuntada en el gráfico.

- Cuadrante 1: Posee una densidad y centralidad elevadas. Los temas situados en él se caracterizan por estar muy desarrollados y ser centrales. Juegan un papel motor dentro del campo de estudio.

- Cuadrante 2: Baja densidad con alta centralidad. Los temas, bien relacionados pero al tiempo poco desarrollados, pueden considerarse como emergentes o como temas puente.

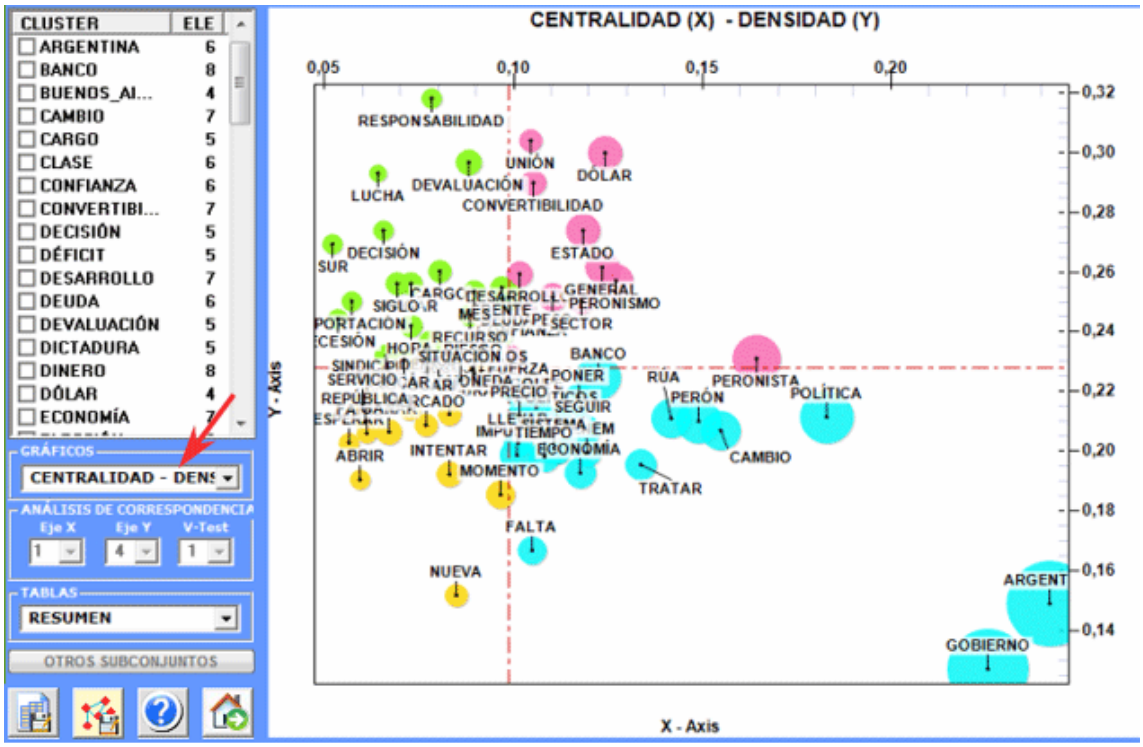
- Cuadrante 3: En él se sitúan los temas muy desarrollados (alta densidad), pero poco centrales (baja centralidad). Estos temas pueden considerarse como altamente especializados y representativos de una alta actividad, pero aislados del centro temáticos.

- Cuadrante 4: La centralidad y la densidad son bajas, por lo que los temas aquí situados poseen un carácter débil y netamente marginal. En este cuadrante suelen aparecer por primera vez los temas y también en muchos casos terminan aquí por desaparecer definitivamente.

En el caso que nos ocupa, la interpretación sería muy similar, pero se representarían términos y sus relaciones, no agrupaciones temáticas.

A continuación se muestra un diagrama estratégico que elabora la herramienta T-LAB[2] con lo que sería la representación de términos más frecuentes en un diagrama estratégico.

[2] Herramienta de análisis de textos. http://tlab.it/es/presentation.php



Ejemplo 2: Diagrama estratégico de T-LAB
Análisis de co-ocurrencias sobre las noticias de la crisis argentina aparecidas en el periódico El Mundo.

Por otro lado, al trabajar con grupos de palabras que son diferentes marcos temporales ya se está dibujando sobre el diagrama estratégico la evolución de las relaciones.

4. Aplicaciones

El análisis temporal de lo que hemos llamado “temperatura del lenguaje y de la cohesión léxica puede hacerse en dos direcciones:

- Hacia atrás: de manera retrospectiva, para tratar de llegar al origen de los tuits o informaciones que han provocado la expansión de una opinión (o rumor). Si se detectan que un grupo de términos aparecen ubicados en el cuadrante 4, que indica que son términos nuevos, se puede ir indagando hacia atrás por sus pares de palabras hasta encontrar un posible origen.

- Hacia delante: para analizar la evolución de un tema en relación a los términos que coocurren. El estudio de esta evolución permite comprobar, por ejemplo, la eficacia de acciones o campañas de imagen o información respecto a un producto o servicio.

El análisis de la cohesión léxica de los textos nos permitirá analizar cómo de fuerte o cohesionado se mantienen ciertos mensajes, y estudiar las causas externas que inciden en este hecho. Puede ser útil buscar la explicación a ventanas temporales en las que los grados de dispersión léxica son muy bajos y al contrario. Esa explicación puede estar relacionada con “acciones” externas, como puede ser alguna acción de marketing o alguna noticia relacionada con el objeto de estudio, por ejemplo. El grado de sincronización de los mensajes puede dejar ver que ciertos movimientos estén dirigidos. Los análisis de coocurrencias desde la perspectiva temporal permiten medir este tipo de cosas y conocer en qué momento se produjo un cambio en la cohesión de los mensajes y poder analizar las causas.

De la misma manera, se puede hacer un seguimiento a diferentes usuarios de la red que se consideren especialmente influyentes en determinados grupos de opinión y analizar tanto con las coocurrencias de términos y el diagrama estratégico, el grado de cohesión de su mensaje a lo largo del tiempo y poder estudiar, con estos indicadores, la influencia que ejerce sobre sus seguidores.

Otra aplicación de este método consiste en comparar métricas entre los mensajes emitidos y las diferentes formas en las que este es recibido y comentado por los usuarios. Está relacionado plenamente con el concepto de intertextualidad comentado en el punto 2 de este trabajo, que determina la manera en que el uso de un cierto texto depende del conocimiento de otros textos. Cuando se transmite un mensaje, el emisor tiene una clara intención comunicativa y esta se representa en la elección intencionada de términos para transmitir una idea. El receptor interpreta ese mensaje y expresa su opinión al respecto. Resulta muy útil comparar las redes de palabras utilizadas para expresar estas opiniones y comprobar si se parecen a las del mensaje original o si los índices son diferentes.

4.1 Ejemplos de uso.

4.1.1 Seguimiento de un rumor.

El análisis de términos coocurrentes y cohesión léxica puede utilizarse para hacer un seguimiento de los temas en Twitter. Estos temas pueden ser rumores. En ocasiones se puede difundir una noticia o información no confirmada, que genera muchas reacciones y se expande rápidamente. Es lo que llamaremos un rumor. Una vez identificada esa información, el análisis de coocurrencias y cohesión léxica puede ayudar en la tarea de identificar el origen del mismo en un análisis que antes hemos llamado "hacia atrás". Igualmente se puede analizar la forma en la que ese rumor se difunde hacia delante en el tiempo.

Para mostrar un ejemplo de este uso, utilizamos las visualizaciones de Twitter Trails[3], una herramienta interactiva, basada en la web que permite a los usuarios investigar las características del origen y propagación de un rumor en Twitter.

Además de tener en cuenta el número de retuits y la influencia de los usuarios que participan de la extensión del rumor, la herramienta analiza la similitud semántica entre los tuits para poder determinar el modo en que el rumor se difunde.

Así pues, tuits con un lenguaje similar se dibujan del mismo color en un intento de visualizar la "independencia del contenido". Una gran variación en los colores indica que hay múltiples fuentes hablando o tuiteando sobre el tema investigado: o bien hay diferentes publicaciones sobre el tema o bien muchos usuarios hablando del tema con diferentes términos. Por el contrario, si encontramos muchos tuits con una escasa variación en los términos empleados, como el caso que presentamos a continuación, indica que la fuente de la información es única y hay muchas posibilidades de que ese rumor sea falso.

Ejemplo 3: Gráfico de propagación de Twittertrails.com sobre una información falsa sobre la llegada de los primeros refugiados sirios a Nueva Orleans

[3] Proyecto de investigación de la Universidad de Wellesley

Como vemos en el ejemplo, la mayoría de tuits son de color verde, lo que indica que hay poca variación léxica en el contenido de los mismos. Esto indica que el mensaje puede ser el titular de un artículo, un único tuit copiado o con pequeñas modificaciones o incluso una única persona difundiendo masivamente un mensaje a modo de spam.

En este trabajo se propone una metodología similar. Una vez establecidos los grupos o marcos temporales del corpus con el que se trabajará, se calcula la distancia léxica (mediante los índices de transformación o intersección, dependiendo del tamaño de la muestra) entre los diferentes grupos. Se calcula la distancia entre cada grupo respecto al anterior. Si el índice el alto (se definirá el grado exigido para considerar dos grupos como lexicalmente cercanos) se asigna el mismo color; si el índice no alcanza el exigido se calcula de nuevo el índice respecto a otro grupo de color. Si no encuentra similitud con otros grupos, se le asigna un nuevo color.

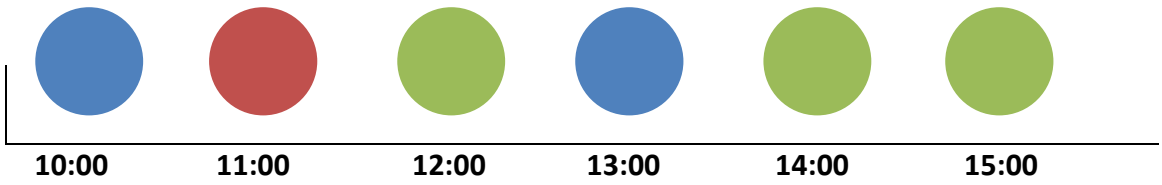


Figura 2: Representación de la asignación de colores según el grado de similitud léxica entre marcos temporales

Esto indica que entre los tuits de la primera y segunda hora apenas hay términos comunes. Los tuits de las 14:00h y las 15:00h tienen muchas palabras en común, con lo que podemos entender que la información contenida en estos tuits es casi idéntica a la de los tuits de las 12:00h. Si el rumor se localiza en el grupo de las 13:00h sería lógico buscar un posible origen en el grupo de las 10:00h.

Cuanto más pequeños sean los grupos o más reducidos sean los marcos temporales, más precisos serán los resultados. El análisis de la cohesión léxica y el estudio de coocurrencias es útil sobre todo para conocer la forma en la que se ha propagado el rumor. Como se ha comentado al principio, este análisis debe completarse con otros datos de tipo cuantitativo como el número de retuits y las relaciones entre usuarios, que dibujarían la línea de propagación.

4.1.2 Aplicación del método en el análisis de un discurso político y las reacciones provocadas en la audiencia

Contamos con el discurso completo. Se hace un análisis de coocurrencias y una tabla de frecuencia de términos.

Además se asume que el discurso, como intervención planeada y estudiada tiene un *timing*, es decir, trata diferentes temas que tienen un tiempo asignado.

Además contamos con el corpus de reacciones obtenidas de la red social Twitter durante 6 horas desde el comienzo de la intervención. Suponiendo que el discurso ha generado muchas reacciones, el marco temporal que se establece es de 15 minutos. Se trabajará pues con 24 grupos o *clusters*.

Posibles análisis:

a. **Análisis por términos.** De cada grupo se obtienen los términos más frecuentes y las coocurrencias que resulten más interesantes. El estudio de un par de coocurrencias en los distintos periodos permite ver la evolución del término y sus relaciones. Se extraen las coocurrencias de cada grupo o marco temporal y se estudia el comportamiento de cada par de palabras en los diferentes grupos.

b. Pero además resulta interesante analizar términos con un comportamiento extraño: puede ocurrir que una palabra con poco peso en el discurso, aparezca con mucha frecuencia en las reacciones. ¿Cómo analizar esto?

    - En primer lugar, la presencia de ese término en su grupo, sitúa al término en un punto en tiempo.
    - A partir de ahí, podemos buscar hacia atrás la aparición del término y comprobar qué textos o influenciadores han podido ser la causa de la explosión del término.

c. Desde el punto de vista del orador, el interés en el análisis de los términos estará en:

    - comprobar qué efecto han tenido determinados temas o palabras.
    - términos más relacionados con el orador durante todo el discurso y las horas posteriores

d. Cálculo del **grado de similitud de los grupos** de reacciones con el discurso original. Se puede calcular el índice de intersección (número de palabras comunes) o el índice de transformación (que tiene en cuenta el total de palabras de cada documento). Debe valorarse la longitud del discurso para decidir si se aplica esta segunda métrica.

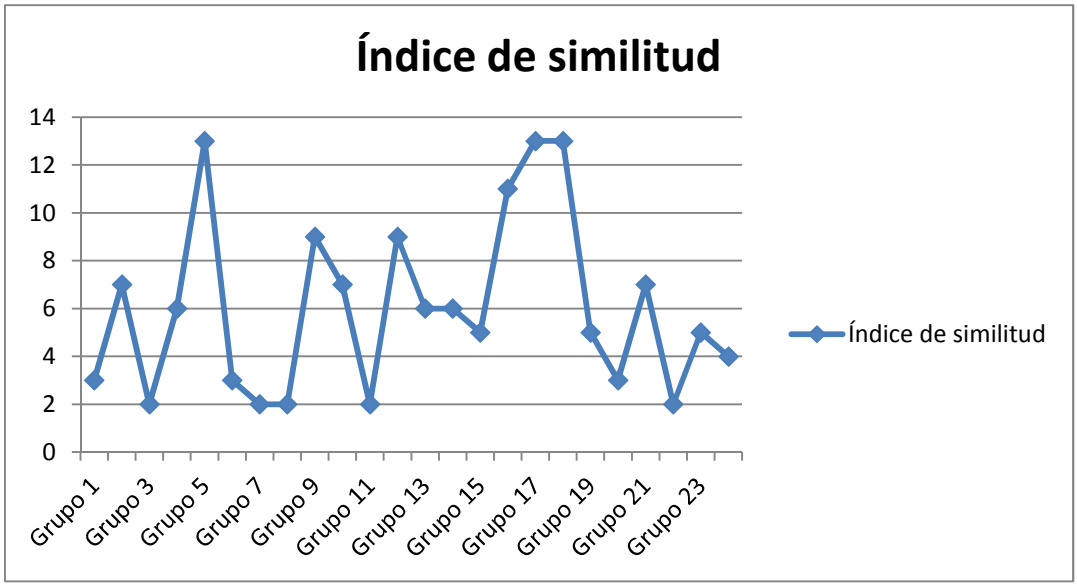


Gráfico 4: Evolución del índice de similitud entre los marcos temporales. Gráfico de elaboración propia

Al compararse el discurso con los grupos temporales, obtenemos estos índices desde el punto de vista cronológico, con lo que se presenta esta información en un gráfico de líneas.

Resulta interesante comprobar si emisor y receptor comparten términos para hablar de lo mismo. Si los índices son altos, indicaría que el receptor está siguiendo con atención el mensaje, si los términos no son los mismos en un alto índice, podría indicar que están apareciendo las interpretaciones del mensaje.

e. Puede ser habitual que en los comentarios del discurso, aunque se hable de los mismos temas, las palabras utilizadas sean bastante diferentes, por las propias interpretaciones o por la adaptación que cada oyente hace a su situación personal. La disposición de los grupos en marcos temporales permite hacer un seguimiento más específico de los términos utilizados por los receptores para cada parte del discurso. Los analistas podrán estudiar mejor las interpretaciones y las diferentes formas de recepción del mensaje lanzado en cada momento del discurso.

f. Densidad y centralidad: para cada grupo de reacciones. **Diagrama estratégico**

Se calcula el grado de cohesión interna de cada grupo para comprobar cómo de fuerte es la relación entre los términos empleados para expresar las reacciones y se analizan en detalle. Además se calcula la relación entre los grupos y se dibuja en el diagrama estratégico.

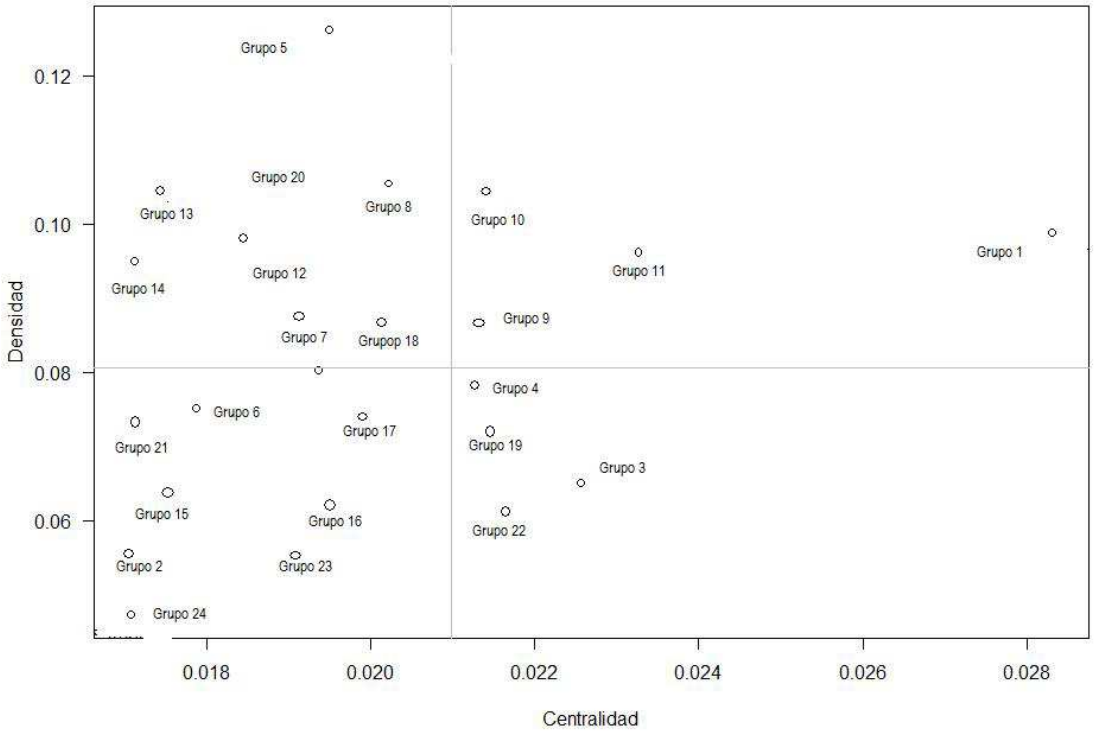


Gráfico 5: Diagrama estratégico. Elaboración propia

Se puede analizar la situación del Grupo 1, que es destacada. Se debe analizar también las situaciones entre cuadrantes opuestos, la situación respecto a ambos niveles y las relaciones ente los grupos.

5. Conclusiones

La aplicación del método de palabras asociadas al análisis de los mensajes de las redes sociales es perfectamente posible. Por un lado, el método aporta la base metodológica de un sistema ya probado y extendido, pero por otro lado, requiere adaptarse al tipo de textos al que va a ser aplicado y al amplio volumen de documentos y la velocidad de procesamiento que requieren.

La principal diferencia, como se ha repetido a lo largo del presente trabajo, consiste en que no se trabaja con palabras clave en el sentido en que lo hace el MPA. Las palabras clave en la indización de los documentos científicos forman parte de un vocabulario controlado y tienen la función de representar al documento para luego facilitar la búsqueda y recuperación. En los análisis de coocurrencias las palabras clave son las palabras más frecuentes. No pretenden representar el contenido, sino reflejarlo.

Por lo tanto se encuentran representados en el diagrama conceptos muy relacionados entre sí, en algunos casos con relaciones de hiperonimia, sinonimia, etc. que en los lenguajes controlados están, precisamente, controladas. Pero es exactamente eso lo que se pretende analizar en este trabajo. Qué palabras se utilizan junto con las de nuestro interés, con qué palabras se relacionan y en qué grado. Es muy importante que sean las propias palabras y sus relaciones las que descubran la información, como se mencionaba en la introducción. La elección de un término u otro puede ser indicativo de información que “a priori” no se habría buscado.

La inclusión de la variable tiempo al estudio del contenido de los mensajes de las redes sociales y en concreto de Twitter, supone un análisis complementario a otros que actualmente se realizan, tanto cuantitativos como cualitativos. Además, de la misma manera que la aplicación del método de las palabras asociadas al tratamiento de la información científica supone la representación del conocimiento de esa temática concreta, su adaptación a los mensajes de Twitter supone también una forma de representación del conocimiento que sobre determinado tema circula por esa red social en el contexto estudiado y sus transformaciones en el tiempo.

6. Referencias

Dlegorreta (10 de Octubre de 2015). ¿Cuánto se puede saber desde los discursos? [Mensaje de un blog]. Recuperado de: https://dlegorreta.wordpress.com/2015/10/10/cuanto-se-puede-saber-desde-los-discursos/

Eíto Brun, R., & Senso, J. A. (2004). Minería textual. *El profesional de la información*, *13*(1).

Escorsa, P., Maspons, R. Módulo 8: La Vigilancia Tecnológica, un requisito indispensable para la innovación. http://docencia.udea.edu.co/ingenieria/semgestionconocimiento/documentos/Mod8IntelgComptInnv.pdf

Fernandez-Amoros, D., Gil, R. H., Somolinos, J. A. C., & Somolinos, C. C. (2010). Automatic word sense disambiguation using cooccurrence and hierarchical information. In *Natural Language Processing and Information Systems* (pp. 60-67). Springer Berlin Heidelberg.

Gamallo, P., Pichel, J. C., Garcia, M., Abuín, J. M., & Pena, T. F. (2014). Análisis morfosintáctico y clasificación de entidades nombradas en un entorno Big Data. *Procesamiento del Lenguaje Natural*, *53*, 17-24.

García, A. G., Ibáñez, A. P., Sapena, A. F., Mancebo, M. F. P., & Moreno, L. M. G. (2015). Herramientas de análisis de datos bibliográficos y construcción de mapas de conocimiento: Bibexcel y Pajek. *BiD: Textos universitaris de biblioteconomia i documentació*, (34), 11.

Gualda, E., & Borrero, J. D. (2015). La'Spanish Revolution'en Twitter (2): Redes de hashtags (#) y actores individuales y colectivos respecto a los desahucios en España. *Redes: revista hispana para el análisis de redes sociales*, *26*(1), 1-22.

Hassan Montero, Y. (2006). Indización social y recuperación de información. *No solo usabilidad*, (5).

Marin, O. (2014). Hacia un método de análisis del lenguaje y contenido emocional en la gestación y explosión del 15M en Twitter. En: *15MP2P. Una mirada transdisciplinar del 15M, pp 331*

Melgar Hinostroza, Luis R (2011). La cohesión textual. UNSCH.

Nazar, R., Vivaldi, J., & Wanner, L. (2012). Cooccurrence graphs applied to taxonomy extraction in scientific and technical corpora. *Procesamiento del lenguaje natural*, *49*, 67-74.

Pericás, J. M. V. (2005). El uso de la teoría de redes sociales en la representación y análisis de textos. De las redes semánticas al análisis de redes textuales. *Empiria. Revista de metodología de ciencias sociales*, (10), 129-150.

Piad-Morffis, A., Estévez-Velarde, S. & Almeida Cruz, Y. (25 a 27 de noviembre de 2015). Extracción de Métricas de Similaridad Semántica de Wikipedia para el Minado de Opinión en Twitter. En Congreso Internacional COMPUMAT 2015. La Habana.

Rey-Vázquez, L. (2009). *Informe APEI sobre vigilancia tecnológica*. Gijón: APEI, Asociación Profesional de Especialistas en Información, 2009..

Ruiz-Baños, R., & Bailón-Moreno, R. (1998). El método de las palabras asociadas (I): la estructura de las redes científicas. *Boletín de la Asociación Andaluza de Bibliotecarios*, *53*, 43-60.

Ruiz-Baños, R., & Bailón-Moreno, R. (1999). El método de las palabras asociadas (II). Los ciclos de vida de los temas de investigación. *Boletín de la Asociación Andaluza de Bibliotecarios*, *54*, 59-71.

Rodríguez, D. H., ; Pardo, C. E. Programación en R del método de las palabras asociadas.

Stuart, Keith; Trelis, Ana Botella. Lingüística de corpus, análisis de redes y matrices de coocurrencias. En *A survey of corpus-based research [Recurso electrónico]*. 2009. p. 612-630.

Valderrábanos, Antonio (11 de abril de 2011). Qué puede hacer la semántica para explotar las redes sociales (y todo internet, ya que estamos). Recuperado de: http://www.marketingdirecto.com/punto-de-vista/la-columna/que-puede-hacer-la-semantica-para-explotar-las-redes-sociales-y-todo-internet-ya-que-estamos-7/