4th International Conference on Industrial Engineering and Industrial Management
XIV Congreso de Ingeniería de Organización
Donostia- San Sebastián , September 8th -10th 2010

Aproximación al estudio de patentes. Indicadores utilizados en la minería de textos

Javier Gavilanes[1], Rosa Maria Rio[1], Ernesto Cilleruelo[2]

[1] Dpto. de Organización de Empresas. Escuela Universitaria de Ingeniería de Vitoria-Gasteiz. Universidad del País Vasco. Calle Nieves Cano 12, 01006. Vitoria-Gasteiz. javier.gavilanes@ehu.es, rosamaria.rio@ehu.es.
[2] Dpto. de Organización de Empresas. Escuela Técnica Superior de Ingeniería de Bilbao. Universidad del País Vasco. Calle Alameda Urquijo s/n, 48013. Bilbao. ernesto.cilleruelo@ehu.es

Resumen

Las bases de datos de patentes son consideradas, por muchos autores, como una fuente de información muy valiosa dentro del proceso de innovación. El presente trabajo estudia cómo la aplicación de la minería de textos al análisis de las patentes proporciona información relacionada con la innovación de la zona a estudio, trasvases de conocimiento, relaciones entre compañías y sectores industriales, determinación de polos tecnológicos o tecnologías emergentes, etc.

Palabras clave: Text mining; Patent information; Patinformatics; Innovation.

1. Antecedentes y Objetivos

El proceso de innovación es considerado por muchos autores como uno de los factores principales para la competitividad de un país o empresa. Dicho proceso no puede ser entendido sin la asimilación, transformación y difusión del conocimiento, por lo que será necesario que las organizaciones cuenten con herramientas que les permitan gestionar y transformar la información en conocimiento útil para la toma de decisiones estratégicas, Escorsa y Maspons (2001).Esto cobra mayor importancia si tenemos en cuenta que nos encontramos en la era del conocimiento, donde a través de Internet, podemos llegar a acceder a ingentes cantidades de información almacenadas en bases de datos, páginas webs…

Las bases de datos de patentes, tanto de pago como gratuitas, son una muy buena fuente de recogida de información, Dou (2004). Ya que estas patentes son el canal principal para dar a conocer los avances científicos, y no pudiéndose encontrar publicada, mucha de esta información, en ningún otro sitio. Los beneficios que aporta el estudio de las patentes en el campo de la Investigación y Desarrollo han sido analizados y demostrados de forma empírica en varios trabajos, Ernst (1998), Acs et al. (2002).

El proceso de obtención de conocimiento a partir de patentes conlleva una serie de etapas que están aglutinadas dentro de lo que podríamos llamar una nueva ciencia, la Patentometría. Uno de los objetivos principales de la Patentometría es proporcionar un acceso eficiente a la información contenida en las patentes y ayudar así a la definición de las estrategias a seguir por las empresas.
Para una primera aproximación a dichas etapas, Trippe (2002) diferencia dos grupos de tareas: las correspondientes a la búsqueda de patentes y las relacionadas con el análisis de las mismas. Más adelante, el mismo autor, Trippe (2003), menciona varias tareas necesarias para el análisis de las patentes: limpieza y agrupación de conceptos,

generación de listas, matrices de concurrencia y gráficos de círculo, realización de clusters con bases de datos estructurados, clusters con bases de datos no estructurados, mapear los clusters, añadir el análisis longitudinal en los mapas, análisis de las citas y las funciones "SAO".

En esta misma línea, Tseng et al. (2007) intentan abarcar el proceso de análisis de patentes con la siguiente secuencia de actividades: identificación de la tarea, búsqueda, segmentación, condensación, realización de clusters, visualización e interpretación de los mismos. Además, proporcionan una serie de técnicas de minería de textos para la realización de cada tarea.

En el artículo de Bonino et al. (2009) se organizan las tareas de la patentometría en tres grupos: búsqueda de patentes, análisis y monitorización de la información. Se discuten los retos y oportunidades que proporcionan los softwares que se utilizan en la patentometría y especialmente, se hace hincapié en buscar mejoras en el análisis semántico.

Por último, comentar el artículo de Moehrle et al. (2010), donde agrupa las tareas en tres bloques: Proceso previo, donde se preparan las patentes para su análisis; Análisis de la Patentes, donde se consigue extraer la información; y Obtención del Conocimiento, donde se visualizan y se evalúan los resultados. En cada una de las tareas, que se pueden realizar de forma automática, se muestra una lista de las herramientas informáticas que lo pueden realizar.

El objetivo de este trabajo es estudiar cómo la aplicación de la minería de textos al análisis del Registro de la Propiedad Industrial (patentes) proporciona información relacionada con la innovación de la zona a estudio, trasvases de conocimiento, relaciones entre compañías y sectores industriales, determinación de polos tecnológicos o tecnologías emergentes, etc.

2. Aplicación de la minería de textos. Estudio de casos

A continuación, se exponen cuatro casos concretos de aplicación del análisis de patentes al estudio de la innovación en diferentes ámbitos como regiones, sectores industriales y/o empresas concretas:

El primer caso es el estudio de la innovación en una región detallado en el artículo de Rio et al. (2008). En él se analiza la evolución del progreso tecnológico en la provincia de Álava a través del estudio de las patentes solicitadas en dicha región. Tras la selección y el filtrado de la información relevante para el estudio, se seleccionan una serie de indicadores para su análisis: solicitantes: porcentaje de particulares y empresas; nacionalidad; número de solicitudes y concesiones. Además se obtuvieron datos a través de la elaboración de diversos cruzamientos: las empresas que más han patentado en el periodo a estudio (1960-2000), número de patentes solicitadas en cada década por cada polígono industrial (Figura 1), discriminación de los diferentes sectores a los que pertenecen las patentes de cada año, CIP (Clasificación Internacional de Patentes) con el campo solicitante graficando el Panorama Industrial Innovador alavés. Por último, para obtener las áreas emergentes de patentabilidad y sus relaciones, cruza nuevamente la CIP con el campo fecha de solicitud (Figura 2). Tras la interpretación de los análisis, se extraen las siguientes conclusiones: La industria alavesa innova como lo demuestran las solicitudes de patentes. Se identifican los principales sectores industriales que potencian la I+D en Álava. Se aprecia cómo a

partir de 1979 hay una disminución de las patentes. Y por último, en la evolución de los polígonos, éstos maduran y disminuyen en gran medida su capacidad innovadora, apareciendo, a partir del 2000, el Parque tecnológico como foco aglutinador de las empresas más innovadoras.

Figura 1. Evolución de la innovación en los Polígonos Industriales Alaveses.

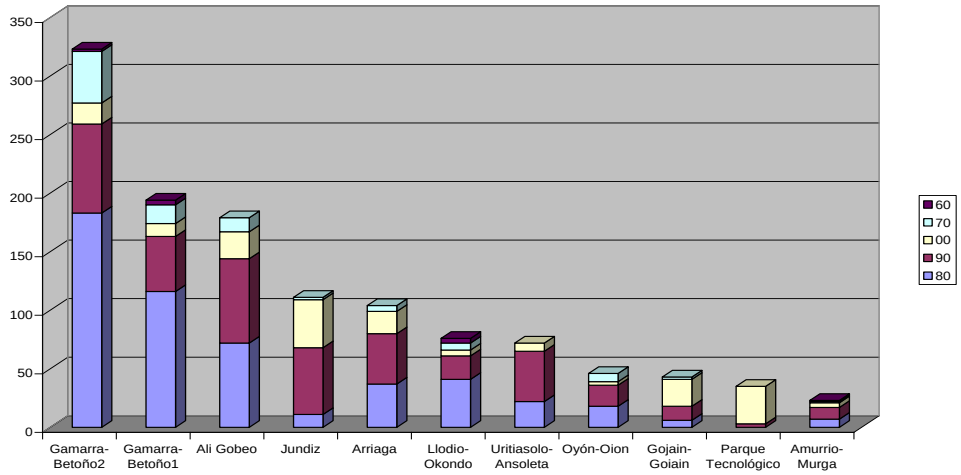
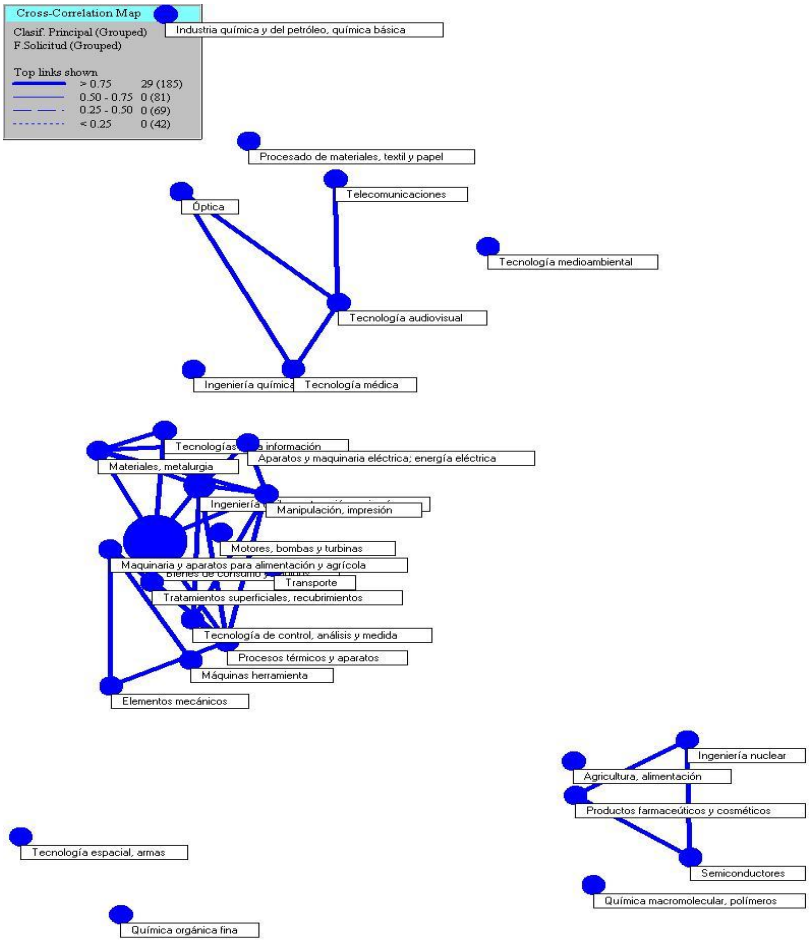


Figura 2. Áreas emergentes de patentabilidad y sus relaciones (CIP- fecha solicitud).



Fattori et al. (2003) aborda el tema del escepticismo en el uso de herramientas informáticas frente a la clasificación manual, para la extracción de información de patentes, por la desconfianza que suscita el no poder controlar la mecánica interna de funcionamiento o "caja negra" de dichos programas informáticos que utilizan algoritmos lingüísticos.
Para ello, analiza, de dos formas diferentes, todas las patentes relacionadas con un sector concreto (empaquetamiento) dentro del periodo 1991-2000. Una, a través un programa informático, PackMOLE™; y otra, a través de la clasificación manual Derwent.
El programa que se utiliza, permite al usuario graduar una serie de parámetros antes de realizar el análisis y agrupar las patentes en clusters y grupos de clusters, como son: determinar el número máximo de clusters; el número mínimo de veces que debe aparecer una palabra clave en un documento para considerarlo relacionado; las "stopwords" o palabras que es conveniente excluir del análisis porque tienden a distorsionar el resultado; y el nivel de semejanza entre dos documentos para incluirlos en el mismo cluster. Una vez realizados los análisis, el programa determina la calidad de los clusters basándose en la homogeneidad intra-cluster y la heterogeneidad inter-cluster. Cruzando este indicador de calidad con los parámetros iniciales y unido al conocimiento del sector que se posea, se puede crear un modelo adecuado para el análisis de las patentes del sector en estudio.
Los resultados del programa fueron revisados manualmente, y se comprobó que el 70% de las patentes estaban agrupadas de forma correcta, considerando el porcentaje como satisfactorio y cuestionando el recelo a utilizar estos programas.
El autor termina concluyendo que la información que se puede conseguir a través de las herramientas informáticas, en este caso el PackMOLE™, es muy superior a la que puede aportar las técnicas clásicas de categorización, clasificación Derwent.

Huang et al. (2003) realiza un estudio para determinar el desarrollo y flujos de conocimiento generados por la investigación en las compañías electrónicas más importantes de Taiwan a través del análisis de citación de patentes.
La muestra final de empresas a estudio se redujo a 58 tras aplicar unos requisitos, como poseer un mínimo de cinco patentes. Tras realizar una revisión de la literatura referente a la información más importante de las empresas seleccionadas, se recogieron de la base de datos americana de patentes (USPTO) las patentes solicitadas por las empresas entre 1998 y 2000 (4.162). Tras el filtrado de los documentos, los indicadores que se estudiaron fueron los siguientes: Número de patentes por empresa, donde aparecen 6 compañías con más de 100 patentes. Número de citas de patentes por cada compañía, el total ascendió a 24.852. Frecuencia de co-citación entre las 58 compañías, se muestran las 10 parejas que con mayor frecuencia (97 veces o más) se citan mutuamente. Coeficiente de correlación, que determinará la distancia entre las compañías en el mapa de clusters; aparecen las 5 parejas con las correlaciones más altas (superiores a 0.98). Mapa de citación de patentes, donde aparece la relación, por cercanía, entre cada una de las compañías, conformándose 6 clusters.
Como conclusiones se menciona cómo el cluster de Semiconductores es el que tiene el coeficiente de correlación más alto; mientras otros, los relacionados con las tecnologías de la información, resultan difíciles de discriminar por la gran interrelación que existe entre ellos.

Bhattacharya S. (2004) estudia el nivel de competencia entre India y China a través de la solicitud de patentes en Estados Unidos. Las patentes en estudio son recogidas de la base de datos de USPTO y categorizadas en NA (Nationally Assigned), patentes

solicitadas por empresas del país en estudio; y NNA (Not Nationally Assigned) las solicitas por otros países diferentes a India o China pero donde su primer investigador es de nacionalidad india o china.
Una vez agrupadas las patentes en estos cuatro grupos, se clasifican por origen: universidades, centros de investigación… (Tabla 1) y por el tipo de patente: "utility", "design" o "plant" (Tabla 2). El grupo más interesante, para ver la evolución tecnológica y el nivel de capacidad competitiva de cada país es "utility", puesto que son patentes basadas en nuevos procesos, mecanismos…Este grupo lo clasifica en 36 sub-categorías o tecnologías.
Tras el análisis, los resultados muestran como tanto en India como China las patentes nacionales como no nacionales van aumentando. China lo hace en mayor medida, y dentro del grupo "utility" y "design". En cambio India solo dentro de "utility". En India las patentes nacionales están concentradas en empresas industriales e institutos de investigación, mientras en China el origen es más variado, jugando las universidades un papel muy importante. En general las patentes de China son más variadas, hay mas colaboración y mas interés de vincular las universidades, industria y centros de investigación, cosa que en India no se aprecia.

Tabla 1. Distribución de las patentes según su origen (India y China).

India – NA & NNA patents						
Period	1996–1997	1996–1997	1998–1999	1998–1999	2000–2001	2000–2001
*Type	NA	NNA	NA	NNA	NA	NNA
Industry	62%	100%	72%	90%	54%	100%
Research institute	29%		10%		20%	
University			3%	2%	4%	
Specialised institute	9%		15%	2%	20%	
Individuals				5%	3%	
China – NA & NNA patents						
Period	1996–1997	1996–1997	1998–1999	1998–1999	2000–2001	2000–2001
Type	NA	NNA	NA	NNA	NA	NNA
Industry	57%	82%	52%	90%	75%	89%
Research institute	8%		16%		9%	
University	20%	15%	16%	3%	7%	4%
Specialised institute	10%		6%		4%	2%
Individuals	5%	3%	10%	7%	5%	5%

Tabla 2. Número de patentes según la actividad (India y China).

India – NA & NNA patents						
Period	1996–1997	1996–1997	1998–1999	1998–1999	2000–2001	2000–2001
Type	NA	NNA	NA	NNA	NA	NNA
Utility	44	40	122	80	206	111
Design	5	2	3	4	0	1
Plant	0	0	1	1	1	0
China – NA & NNA patents						
Period	1996–1997	1996–1997	1998–1999	1998–1999	2000–2001	2000–2001
Type	NA	NNA	NA	NNA	NA	NNA
Utility	53	77	99	118	133	272
Design	8	6	11	28	148	133
Plant	0	0	0	0	0	0

3. Conclusiones

Tras el análisis de los casos podemos concluir que el análisis de las patentes de un área geográfica permite identificar sus motores de innovación. Su estudio permite visualizar la evolución de los indicadores de innovación a lo largo del tiempo y la visualización de los campos clave permite comprender el funcionamiento de las redes de innovación.

Para el análisis de las patentes, son muchas las herramientas que existen en el mercado, y prácticamente, todos los autores mencionados en este artículo, coinciden en que no existe una única herramienta que realice de forma altamente eficiente cada una de las tareas necesarias para extraer información valiosa de dichas patentes. El tipo de usuario que esté realizando el análisis y el campo de estudio en el que se esté investigando, son algunos de los aspectos que determinarán la correcta elección de las herramientas a utilizar.

A la hora de estudiar los flujos de innovación de un grupo de empresas, sean éstas seleccionadas por ubicación física o por pertenencia a un sector concreto, se tienen varios indicadores que aportarán la información buscada: Número de patentes por compañía; Número de patentes por sector; Número de citas entre patentes de las compañías a estudio (co-citación), para determinar las cooperaciones que existen; y cualquier cruzamiento de datos que se considere adecuado para extraer nueva información, como puede ser el sector al que pertenece la patente y su fecha de solicitud, con lo que se extraerán las tecnologías emergentes.

Los autores coinciden en que la interpretación correcta de los resultados conlleva un conocimiento extra de las empresas en estudio, ya que ciertos indicadores como las co-citaciones, pueden estar influenciadas por otro tipo de intereses más allá de los puramente científicos.

Los indicadores que se muestran en los casos, como las recomendaciones a la hora de seleccionar las muestras, filtrarlas, analizarlas e interpretarlas son útiles para el estudio de las patentes del País Vasco.

Referencias

Acs Z.J.; Anselin L.; Varga A. (2002). Patents and innovation counts as measures of regional production of new knowledge. Res Policy, Vol. 31, pp. 1069–1085.

Bhattacharya S. (2004). Mapping inventive activity and technological change through patent análisis: A case study of India and China. Scientometrics, Vol. 61, No. 3, pp. 361-381.

Bonino D.; Ciaramella A.; Corno F. (2009). Review of the estate-of-the-art in patent information and forthcoming evolutions in intelligent patent informatics. World Patent Inf; in press. Doi: 10.1016/j.wpi.2009.05.008.

Dou H.J.M. (2004). Benchmarking R&D and companies through patent analysis using free databases and special software: a tool to improve thinking. World Patent Information, Vol. 26, pp. 297-309.

Ernst H. (1998). Patent portfolios for strategic R&D planning. Eng Technol Manage, Vol. 15, pp. 279–308.

Escorsa P.; Maspons R. (2001). De la vigilancia tecnológica a la inteligencia competitiva. Pearson Educación, s.a.

Fattori M.; Pedrazzi G.; Turra R. (2003) Text mining applied to patent mapping: a practical business case. World Patent Inform, Vol. 25, No. 4, pp. 335–342.

Huang M.H.; Chiang L-Y.; Chen D.Z. (2003). Constructing a patent citation map using bibliographic coupling: A study of Taiwan´s high-tech companies. Scientometrics, Vol. 58, No. 3, pp. 489–506.

Moehrle M.G. et al. (2010). Patinformatics as a business process: A guideline through patent research tasks and tools. World Patent Information, doi: 10.1016/j.wpi.2009.11.003.

Rio R.M. et al. (2008). PatentAlava. Dinamics of innovation strategies and their relationship with the evolution of patents. The Alava province case. 5th International Scientific Conference Business and Management' 2008, pp. 475–479.

Trippe, A.J. (2002). Patinformatics: identifying haystacks from space. Searcher, Vol. 10, No. 9.

Trippe, A.J. (2003). Patinformatics: tasks to tools. World Patent Inform, Vol. 25, No. 3, pp. 211–221.

Tseng Y.H.; Lin C.J.; Lin Y.I. (2007). Text mining techniques for patent analysis. Information Processing & Management, Vol. 43, pp. 1216–1247.