

Research-based knowledge to corporate businesses: the Danish experience

La expansión del conocimiento basado en la investigación a las empresas corporativas: la experiencia danesa

Anders Mejlbjerg*

Abstract

One of the goals in the Danish Government's strategy for growth in Denmark is to promote the awareness, access and use of research-based knowledge to citizens and institutions, both private and public. In gaining better use of research-based knowledge, Denmark's chances of being able to play a prominent role as a knowledge society, both nationally and internationally, will substantially increase. To transform the strategy into practice Denmark's Electronic Research Library, DEFF, has been created. In this article business consultant of DEFF, Anders Mejlbjerg, describes how DEFF is working to expand the awareness of research-based knowledge to various public institutions and to small and medium-sized private enterprises, known as SMES.

KEYWORDS: SME, Denmark, research-based, strategy, healthcare.

RESUMEN

Uno de los objetivos de la estrategia del gobierno danés para el crecimiento en Dinamarca es promover el conocimiento, acceso y uso del conocimiento basado en la investigación a los ciudadanos y las instituciones, tanto privadas como públicas. Al hacer mejor uso del conocimiento basado en la investigación, las posibilidades de Dinamarca para jugar un papel destacado como una sociedad del conocimiento, tanto a nivel nacional como internacional, aumentarán sustancialmente. Para transformar la estrategia en la práctica, se creó la Biblioteca Electrónica de Investigación de Dinamarca, DEFF. En este artículo, el consultor comercial de la DEFF, Anders Mejlbjerg, describe cómo la DEFF está trabajando para ampliar la concientización del conocimiento basado en la investigación a diversas instituciones públicas y para las pequeñas y medianas empresas privadas, también conocidas como PYMES.

PALABRAS CLAVE: PYMES, Dinamarca, investigación, estrategia, sector salud.

* Business consultant, former trade publisher / Danish Agency for Culture, DEFF / H.C. Andersens Boulevard 2, 1553 Copenhagen V, Denmark. Correo electrónico: anm@kulturstyrelsen.dk

This article intends to give the readers a brief insight into the work we do in the Danish Agency for Culture's digital office. In particular what we are trying to achieve is to link and relate to the research-based institutions, libraries and –in some cases– corporate businesses in Denmark.

The Agency for Culture is an agency under the Ministry of Culture, which means that we are part of the public administration in Denmark. The Agency for Culture employs 260 people throughout the agencies in 15 highly specialized divisions.

The agency carries out the cultural policies of the Danish government within the visual and performing arts, music, literature, museums, historical and cultural heritage, broadcasting, libraries and all types of printed and electronic media. The agency works internationally in all fields, and the projection of Danish arts and cultural life across the border and out to other countries is a top priority.

I work in a subdivision of the digital office named DEFF. The four letters stand for the Danish abbreviation of Denmark's Electronic Research Library - and for convenience, I will be using the Danish abbreviation, DEFF.

My office is a bit like a shop-in-shop. It is formed on a business-model that doesn't really have an appropriate English term. The closest I can get as an explanation of this in English is under the heading "Commercial Activities". This is slightly misleading, as we do not conduct business of a traditional commercial nature but primarily deals with, and tend to, customers and companies that are either fully or partially public, or - in the case of private companies - are conducting research in separate divisions within the companies. The licenses that we are handling do not allow the users to make use of them in a straightforward commercial way; they have to have some sort of research involved in the use of the licenses.

Denmark is a very small country. We are located on the top of the European continent and at the bottom of the Scandinavian Peninsula. The population of Denmark is 5,5 million people, which is somewhere between a third and a quarter of the population of Mexico City.

Denmark is often named as one of the world's happiest countries. The American television network CNN has tried to find the answer to what makes Denmark a happy country, and after a visit to one of our largest hospitals CNN concluded that it must be because of the fact that in Denmark it is free for the public to be treated in a hospital and is also free, while hospitalized. It may well be that this is one of the reasons; it certainly cannot be because of the weather or the high cost of living ... But anyway, I have just brought this example to your attention due to the fact that the Danish hospital system is one of the largest stakeholders in the organisation I work for.

Denmark has never been a great industrial country with lots of resources. We have a handful of large international companies whose export contributes to the economy of the country (names like Lego, Novo Nordisk, Danfoss and Maersk might sound familiar to you) - and agriculture in Denmark is still quit a large contributor. But beyond that, we must primarily create our living as a knowledge society. This is particularly important to support this condition, therefrom DEFF was born.

And to understand the work of my office within DEFF, you need to know a little bit about DEFF in general: DEFF's duties are wide ranging, but our primary and most important task is through our work to contribute to the most effective and optimal use of research-based information in Denmark.

This is partly done through a license-team engaged in the procurement of electronic licenses to academic and research libraries and institutions - both public and private. I am located in the license-team where I, along with my colleagues, are working to negotiate licensing agreements containing the most relevant content at the best possible conditions for the lowest possible rates for the participating institutions. DEFF's license area also handles the administration of the licensing agreements, including billing, renewal, IP administration, archiving, access to relevant contractual documents and help-desk tasks in relation to the institutions.

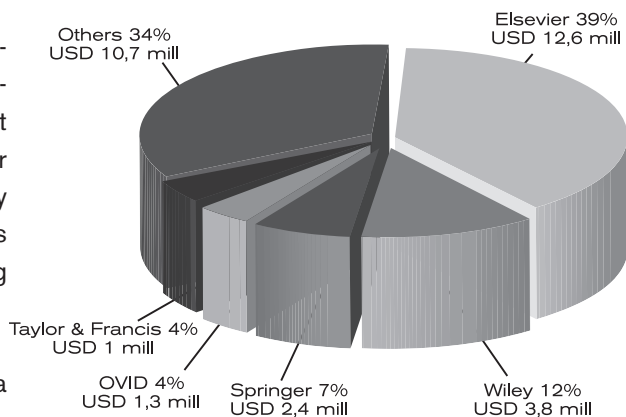
DEFF is owned jointly by the Ministry of Higher Education and Science, the Ministry of Education and the Ministry

of Culture. Our overall mandate is written in the Danish Finance Act, where it says that: "DEFF is to provide support for joint development projects, infrastructure development and to manage the joint purchase of licenses."

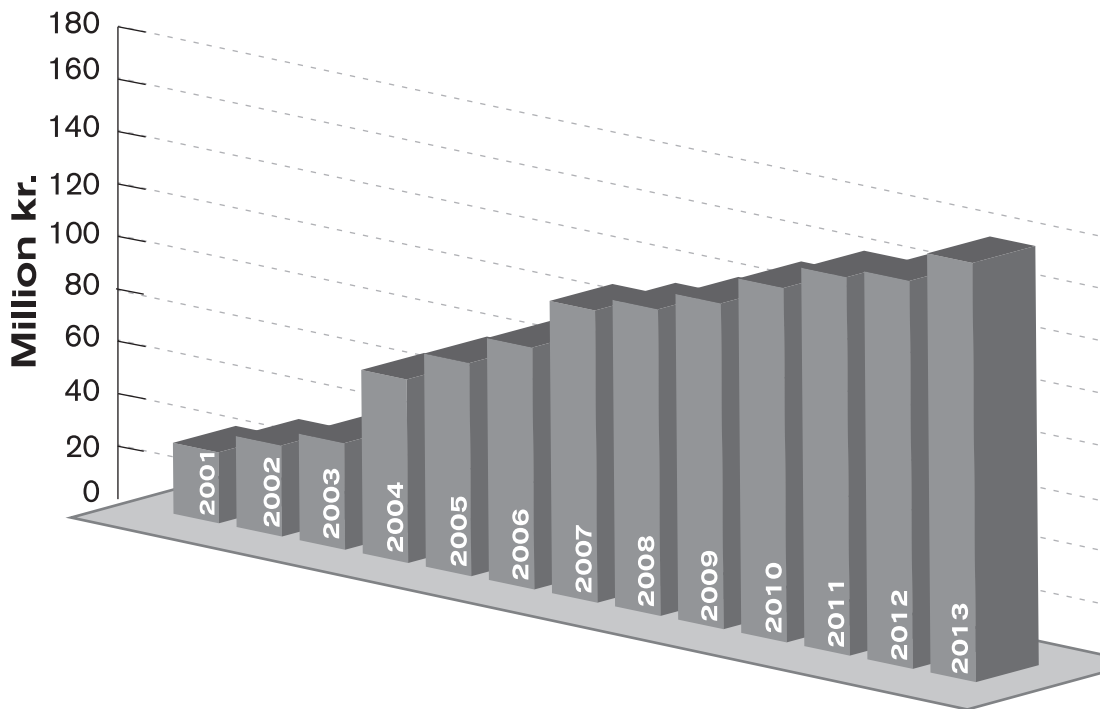
So, as you can see it is the Danish government's desire to have DEFF negotiate and conclude contracts for electronic licenses on the behalf of the research libraries. This joint buying of licenses is cost neutral and without additional cost to the institutions that fall within one of the three owning ministries. The only thing that DEFF invoice the libraries who wish to participate in such cooperation is the pure cost of the licenses as negotiated with the publishers.

The Agency of Culture is the hosting institution of DEFF, and the operational responsibility for licensing negotiations is therefore placed at the Agency.

The DEFF licensing team currently manages nearly 200 different contracts with about 80 publishers for nearly 225 different institutions. The contracts mainly cover product-content such as databases, journals and e-books. Our yearly turnover is 175 million DKK, roughly 32 million USD. Approximately 60% of the turnover comes from the three major suppliers in the market: Elsevier / Springer / Wiley, while the remaining 40% is made up of a number of other smaller suppliers.



The development in the turnover in the DEFF's license area 2001-2013 is summarized on this figure:



This raising curve reflects several factors, including the addition of new institutions and customers, and an increase in the amount of digital resources purchased. But it also reflects a great increase on the costs of the licenses annually imposed by suppliers.

As mentioned earlier, DEFF is owned by the Ministry of Higher Education and Science, the Ministry of Education and the Ministry of Culture and only institutions belonging to these three ministries are eligible to receive DEFF's services in license negotiation and license management, free of charge.

DEFF was founded in 1996, and historically institutions not belonging to the three owning ministries had also been allowed to join the consortium at no charge - including several private companies and institutions.

In 2011 the steering committee behind DEFF decided that the number of institutions outside the ownership group had reached a level that called for an alternative solution for these. Moreover, it was necessary, according to Danish law, that DEFF, as a public authority, not to act anti-competitively in relation to the increasing number of private clients; meaning that we at DEFF were acting against the law if we were to offer a service for free that usually costs money at private providers of licensing.

The steering committee decided to establish a business model in which institutions, in addition to the actual price of the purchased licenses, are charged a small administration fee calculated as a percentage of their purchases (typically 3-5% depending on size).

Given the fact that we are not allowed to take advantage of our governmental status and thus "outcompete" the private vendors, we also had to construct a model very similar to the one used by others in the market.

A consulting firm was asked to prepare a framework for the management of these customers based on the Danish State's rulings on Commercial Activities. After a thorough analysis, the Commercial Activities division was established in 2012 and a full-time business consultant was hired to operate the new system. In 2013 I joined in this

position, and, today, I am running the department on my own, with the help of an external service partner.

When the new scheme was set up, the customers affected had to be transferred to the new system. Despite criticism from a number of the customers that they now had to pay for previously free services, we succeeded in getting 96% of the initial customer-base to join the new system in the Commercial Activities division, and today we have a turnover of around 26 million. kr., and revenue of about 1.5 million. kr. annually. The customers are expressing great satisfaction with the services they receive from Commercial Activities and most recently our "key account", The Danish Regions, which represents 75% of the annual turnover (about 20 million. Kr.) renewed their contract with us.

This governmental business model dictates that, throughout a 4 year period we have to be in balance financially; you can have "fluctuations" over the period, but by the end of the 4 year period - which is by the end of next year - the Commercial Activities division has to present its owning ministries with "healthy" black figures in the financial statement in order to extend the period for another 4 years.

The earnings in the Commercial Activities division derive only from administrative expenses added to the price of the license, and a start-up fee for new customers. The payment from the customers is split into two payments, one in May and one in October - allowing the customers to take advantage of an extended credit. We can offer that as state consortium, since we do not pay interest on the money that we are pre-paying to the publishers on behalf of the customers.

As mentioned earlier, our biggest customer is the Danish Regions. The Regions are the administrative and operational owners of the Danish hospitals. We have 30 public hospitals in Denmark as of July 2014, compared with 37 in July 2011. The trend is pointing towards fewer and larger hospitals, which most likely will create some challenges in the way the future licensing negotiations will be conducted with publishers. Many of them have a tendency of basing their licensing-pric-

es on the number of hospital units rather than pricing them based on the number of relevant users. I doubt that publishers will decrease their pricing, though there are to be fewer physical units of hospitals in Denmark in the years to come... most likely the publishers will find a way to continue to make money, including the licenses that they are selling to the Danish hospitals, however few they might be in the future.

The hospitals are evenly spread across all five regions, with the highest number in the Capital Region. A large group of hospitals are affiliated with one of the country's four major universities in the cities of Aalborg, Aarhus, Odense and Copenhagen - and consequently intensive research is taken place in these hospitals. That is also the case for the best part of the other hospitals as well, and throughout the country health professionals and students need to continually access the latest and most current knowledge in their specific fields.

To obtain this knowledge the Danish Regions created in 2007 a region-license-group, and this group chose to buy their licenses through DEFF. Because the Danish Regions does not belong to either one of the founding ministries of DEFF, as previously explained, the responsibility for the administration of their license-portfolio belongs in my department. The region-license-group consists of representatives from all the five Danish regions, 2-3 from each region, and it is mainly librarians from hospitals participating in the group. They consult on the various needs with medical staff at each hospital, and subsequently address me with what licenses are required. It is my job to obtain licenses on the best possible terms for the best price possible, and to play the role as a secretary of the group.

The regions have an annual licensing budget of approximately 22 million DKK, roughly 4 million USD, and they typically buy license-packages from major publishers such as Springer, Wiley, Oxford University Press and Elsevier. The DEFF-consortium has already agreements with these, and many more, publishers. Therefore I generally do not buy separate licenses to the hospitals, but instead I use the existing DEFF agreements with the major publishers. There are a few exceptions; for example, only the Danish hospitals have purchased the product

“Clinical Key” from Elsevier, so that the contract was dealt and handled by me on behalf of the hospitals.

In addition to the nationwide licenses for the hospitals, the individual hospitals also purchase a number of licenses through my office. It is typically about individual journals, additional packages etc. These agreements are negotiated by me, but they are only possible if the DEFF-consortium already has an agreement with the given publisher. We are happy to take on contracts with new publishers on request, but not all publishers are willing to offer a consortium-discount – and without any discount we cannot offer our customers any added value, therefore we refuse to take on the new account or publisher.

In Commercial Activities we currently have 39 customers. They roughly fall into two groups; institutions within health-science and “the rest”.

In our portfolio there are also, as mentioned earlier, a number of private companies. They are obviously in the market to make money, but since most of the licenses we purchase in the consortium are based on academic and research-based knowledge to higher education or innovative development in non-profit organizations and not for commercial use in the classical sense, is the only the research departments in the private companies I am dealing with. Carlsberg is an example of this; I operate their research department, while the rest of Carlsberg purchases a number of additional licenses through a private vendor in Denmark.

Given that we are a public institution, we are not allowed in any way to outperform the private players in the market for licenses. That also means that we are unable directly to seek out and “steal” customers from the other vendors. However, if a given customer gets in touch with us at their own account, we then of course welcome the request and try to add the customer into our portfolio.

Growth through innovation on the basis of licensed scientific and research based information is an important part of DEFF's strategy - and that is why we are working continuously to expand this.



Currently, we have several projects in the Commercial Activities division: Along with Springer Verlag, we have partnered with Danish Industries, an association of industrial and other companies in Denmark, to invite a number of small and medium-sized enterprises to access the content in Springer's R & D package. Access is made available as a free trial for a four-month period in which the use will be monitored in order to let the registered company use the statistics as a contributing factor to help them decide whether to continue after the end of the trial period.

If they wish to do so they have to pay for the access at a Pay-per-Use Deposit Account set up by Springer in association with my office; the company buys a package with a predetermined number of downloads that may be renewed or extended as they are being used. The goal is for Springer to get new customers, for Danish Industry to act as facilitator for their members, and for DEFF to meet the Danish government's strategy to

make research-based knowledge available to small and medium-sized enterprises for the benefit of innovation throughout Denmark.

The JISC Collection in England is currently working on a similar project, but as far as I am aware, they have experienced quite some problems with the technical set-up between the 4 participating publishers and the chosen companies within the pilot project.

However, we at DEFF in Denmark, through the organization "Knowledge Exchange", are having regular talks and meeting with consortium-colleagues in The Netherlands, Germany, Finland and England, hopefully we can benefit from each other's experiences in this and other fields.

Over the last couple of months, the Commercial Activities division has received more and more inquiries from individual municipalities who would like to gain access

to the same health-care knowledge as the Danish Regions are gaining through us. To follow up on these inquiries we have started negotiations with the main organization of the Danish municipalities, in order to investigate if a business model can be made to offer national licenses for all of Denmark's 98 municipalities.

We also have a project in the pipeline with the provider of press clippings, Infomedia. We will try to put together a package that sells relevant press clippings in addition to a license purchase - as a kind of pay-per-view. A lot of small and medium-sized businesses pay Infomedia to monitor relevant press-clippings, but these companies do not necessarily have neither the time nor the opportunity to seek out relevant research based knowledge. If we can find a way to coordinate their press-clippings with access to various other research based content aimed at their specific fields of interest, it could hopefully help to raise awareness of how much scientific knowledge is actually available at a reasonable cost – just waiting to be read and used by the companies in their daily work.

In parallel with the activities in the Commercial Activities division, work is being made in the steering committee for DEFF to facilitate a number of major projects. Many of them are aimed at investigating different solutions as to how the use of research-based knowledge for innovation and growth can be further optimized.

For instance a project called "Borrow a librarian": It was designed to test a specific model of how research librarians could be an active player in the development of new business. In the project a librarian from a research library visited selected companies, who then were taught the ways of knowledge discovery, along with getting help in solving a series of practical knowledge discovery tasks. During the visits, the librarians' understanding of corporate work increased, while the companies were made aware of the value of the services and expertise libraries can offer.

Unfortunately this project is no longer active, and the effects of it have been difficult to assess.

This is a good example of the fact that many companies and businesses in the private sector often find it difficult to spend the required amount of time to let the projects live on after we, as initiators and promoters, have pulled out of the projects.

But we continue undaunted, to meet the Danish government's innovation strategy and in the hope that practice makes perfect. ☘

Consulted resources

Danish Regions [en línea]. Danske Regioner. <<http://www.regioner.dk/in+english>> [Consulta: septiembre, 2014].

Deff [en línea]: *Danmarks Elektroniske Fag- og Forskningsbibliotek*. <<http://www.deff.dk/>> [Consulta: septiembre, 2014].

Deff [en línea]: *Denmark's Electronic Research Library*. <<http://www.deff.dk/english/>> [Consulta: septiembre, 2014].

Regionerne er garanteret for et sundhedsvæsen af høj kvalitet [en línea]. Danske Regioner. <<http://www.regioner.dk>> [Consulta: septiembre, 2014].

Minería de texto: una visión actual

Text mining: a current view

Marcial Contreras Barrera*

RESUMEN

En la actualidad, debido a las actividades sociales, económicas y académicas, en conjunto con la utilización de las Tecnologías de Información y Comunicación (TIC), la producción de conocimiento y de información digital se genera cada vez con mayor rapidez y, en ocasiones, se acumula en enormes volúmenes de información en espera de un procesamiento adecuado. Áreas como la minería de datos y la minería de texto son utilizadas en el procesamiento automatizado de información, las cuales ofrecen la posibilidad de explorar grandes cantidades de datos o texto, estructurados y no estructurados, para buscar patrones, encontrar relaciones y extraer conocimiento; por lo que el propósito del presente artículo es mostrar las características y aplicación de la minería de texto como área emergente en el procesamiento de información digital.

Palabras clave: Minería de texto, información digital, procesamiento.

Abstract

Nowadays, due to the social, economic and academic activities, together with the use of information and communication technologies (ICT), knowledge production and digital information is generated increasingly faster and sometimes accumulates in huge volumes of information, pending for proper processing. Areas such as data mining and text mining are used in the automated processing of information, which provide the ability to scan large amounts of data or text, structured and unstructured, to look for patterns, find relationships and extract knowledge. The purpose of this article is to show the features and applications of text mining as an emerging area of digital information processing.

Keywords: Text mining, digital information, information processing

* Departamento de Producción, Dirección General de Bibliotecas (DGB). Universidad Nacional Autónoma de México (UNAM). Biblioteca Central, Ciudad Universitaria, 04510, México, D. F., México. Correo electrónico: marcial@dgb.unam.mx

Introducción

La producción e incremento del volumen de información digital en los últimos años ha sido de forma exponencial, de tal suerte que en la actualidad es necesario contar con equipos de cómputo de alto rendimiento con capacidad de almacenamiento, desde varios gigabytes hasta varios hexabytes¹ de información, y al mismo tiempo con la infraestructura de comunicaciones adecuada, anchos de banda del orden de los Mbits/s hasta los Gbits/s²; ejemplo de ello son los estudios realizados por la empresa EMC³, en donde muestra que la información digital mundial se duplica cada dos años. Para el año 2011 la empresa tenía calculado un volumen de información mundial de 1,8 zettabytes⁴ y estima que para 2020 el mundo va a generar 50 veces la cantidad de información y 75 veces el número de “contenedores de información”, mientras que el personal de TI (tecnología de información) para gestionar crecerá menos de 1.5 veces.

La mayoría de los datos que recopilan, crean y gestionan las compañías hoy en día están desestructurados; repartidos en documentos de procesadores de texto, hojas de cálculo, imágenes y vídeos que no se pueden interpretar fácilmente⁵ por lo que es necesario el procesamiento automatizado de éstos.

En lo referente a la producción de documentos científicos, según las estadísticas de MEDLINE se estima que existen alrededor de 13 000 a 14 000 títulos biomédicos publicados actualmente en todo el mundo de los cuales 5 300 títulos están indexados y se incluyen en la base de datos MEDLINE.⁶

Un área que ha experimentado este incremento exponencial ha sido el de la producción de información,⁷ que junto con las tecnologías de información han dado lugar a complejos sistemas de gestión y provisión de información para diversas tareas que se desarrollan en la sociedad contemporánea. En especial, la educación superior precisa contar con eficaces herramientas de información para la enseñanza, el aprendizaje y la investigación.

A partir de la segunda mitad del siglo XX se inicia la automatización de información dando lugar a la creación de bases de datos de tipo referencial, que proporcionaban a los usuarios datos básicos de un documento como el título, autor, la fuente, palabras clave, entre otros. Con el desarrollo de la informática mejoran las técnicas de indexación, organización y clasificación y surgen nuevos servicios y documentos electrónicos tales como libros, revistas, periódicos, tesis, directorios, diccionarios electrónicos y bases de datos de texto completo que permiten a los usuarios la obtención automatizada del documento.

Durante más de una década, Don Swanson ha argumentado que es posible recuperar nueva información que se deriva de colecciones de textos, debido a que los expertos sólo leen una pequeña parte de lo que se publica en sus áreas de conocimiento y no son conscientes de la evolución de los campos relacionados.⁸ Por lo tanto, debería ser posible encontrar los vínculos entre la información útil en la literatura de las diferentes áreas del conocimiento.

Minería de texto

La proliferación del uso de dispositivos computacionales y de comunicación para la producción de información

¹ 10¹⁸ bytes = 1 millón de millones de bytes

² Transmisión de bits por segundo, 106 bit = 1 000 000 bit/s = 1 Mbit/s (un megabit o un millón de bits)

³ *Digital Universe* [en línea]. EMC2. <<http://www.emc.com/leadership/programs/digital-universe.htm>>

⁴ Un *zettabyte* es una unidad de almacenamiento de información cuyo símbolo es el ZB, equivale a 10²¹ bytes.

⁵ BEATH, Cynthia, Becerra-Fernández, Irma, Ross, Jean, Short, James. El valor de la explosión de la información [en línea]. *IQ Intelligence quarterly: diario de análisis avanzado*, 2013, p. 8-11. <http://www.sas.com/offices/europe/spain/IQ/Q113/IQBigAnalytics_GrandesOportunidades.pdf>

⁶ *Journal Selection for medline Indexing at NLM* [en línea]. U.S. National Library of Medicine. <http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a15>

⁷ Documentos, reportes, e-mails y páginas web.

⁸ SWANSON, Don R. Complementary structures in disjoint science literatures.

digital, y en particular en la producción de documentos textuales, ha generado la necesidad de desarrollar métodos, algoritmos y sistemas capaces de realizar el procesamiento automatizado de datos textuales estructurados, semi-estructurados y no estructurados para su organización y consulta, y con ello el surgimiento de áreas de estudio de la información como la minería de texto.

La minería de texto es un área de investigación del procesamiento automático de la información. Se define como el proceso de descubrimiento de patrones

Finalmente, la minería de texto se puede ver como un área que se encarga del estudio de la información digital y en particular de los documentos textuales, con el objetivo de descubrir tendencias, patrones, desviaciones y asociaciones de una colección de textos, para –en última instancia– pasar al descubrimiento de conocimiento en considerables cantidades de información no estructurada. Como ejemplo de lo anterior, la figura 1 muestra la relación que existe entre la insulina y la vasculitis

Leukocytoclastic vasculitis associated with insulin aspart in a patient with type 2 diabetes. »XML

S Marusic, V Vlahovic-Palceviski, D Ljubanovic, pp. 603-5, Volume 47, Issue 10, International journal of clinical pharmacology and therapeutics, 2009 [PMID:19825323]

OBJECTIVE: To report a case of **leukocytoclastic vasculitis** associated with **insulin** aspart therapy. CASE SUMMARY: A 56-year-old man was admitted to the Department of Endocrinology because of a poorly controlled **Type 2 diabetes**. In an attempt to reach a tight blood glucose control, an intensive **diabetes** management consisting of one evening dose of intermediate-acting **NPH insulin** and three preprandial doses of short-acting **insulin** aspart was introduced. Two weeks following **insulin** aspart introduction the patient developed palpable **purpura** on distal parts of the upper and lower limbs. Four days after the onset of **purpura**, a **skin** biopsy was performed. Histological examination showed **vasculitis** with perivascular infiltrates of lymphocytes and erythrocyte extravasation. Direct immunofluorescence was negative. On the day the purpuric eruptions appeared, **insulin** aspart was substituted with regular **human insulin**. All **skin** lesions disappeared spontaneously within 8 days. **Insulin** aspart was not re-administered. DISCUSSION: Other possible causes of **vasculitis** in this case were excluded by diagnostic tests. **The temporal relationship between the insulin aspart administration and the occurrence of purpura**, with no further episodes of **skin** eruptions after discontinuation of the drug, support the hypothesis of an **insulin aspart** caused **vasculitis**. Based on the Naranjo's algorithm, the adverse drug reaction could be considered possible. CONCLUSION: Clinicians should be aware of the possibility of **leukocytoclastic vasculitis** occurring during **insulin** aspart treatment. »XML

Figura 1: Relación entre insulina y vasculitis.

interesantes y nuevos conocimientos en una colección de textos, es decir, es el proceso encargado del descubrimiento de conocimientos que no existían explícitamente en ningún texto de la colección, pero que surgen de relacionar el contenido de varios de ellos.⁹ Para Witten, la minería de texto es el proceso de analizar escritos o conjuntos de enunciados para extraer información que resulta útil para propósitos particulares.¹⁰ Según Sukanya, la minería de texto es un campo interdisciplinario joven el cual se basa en la recuperación de información, minería de datos, aprendizaje de máquina, estadística y lingüística computacional.¹¹

Para Sánchez¹² la minería de texto hace referencia al descubrimiento no trivial potencialmente útil de conocimiento partiendo de una colección de documentos de texto no estructurado. Y puede realizarse una analogía con la minería de datos, encargada de descubrir conocimiento en bases de datos.

Para el logro de los objetivos la minería de texto utiliza diversos métodos y tecnologías como: la recuperación de información, métodos estadísticos y matemáticos, procesamiento de lenguaje natural, métodos de clasificación y agrupamiento de datos, entre otros.

En las instituciones bibliotecarias o unidades de información, las técnicas de minería de texto tienen mucho que ofrecer a las bibliotecas digitales y a sus usuarios, por esta razón existen trabajos desarrollados para integrar las técnicas de minería de texto a los sistemas encargados de gestionar las colecciones de las bibliotecas digitales, por ejemplo el sistema de cómputo Greenstone utilizado en la gestión de bibliotecas digitales desarrollado en Nueva Zelanda.

Greenstone organiza los documentos y los pone disponibles en Internet; además de permitir la búsqueda en texto completo cuenta con módulos para realizar la

⁹ SWANSON, Don R., *idem*.

¹⁰ Text mining in a digital library. Ian H. Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan.

¹¹ SUKANYA, M., Biruntha, S. Techniques on Text Mining.

¹² SÁNCHEZ, D., MARTÍN-BAUTISTA, M. *Un enfoque deductivo para la minería de texto* [en línea]. <<http://www.softcomputing.es/estylf08/es/2006-XIII%20Congreso/articulos/40.pdf>>

minería de texto a través de la extracción de acrónimos y su definición a partir del texto completo y de la colección, agregando esta información a su base de datos como metadatos. El sistema Greenstone contiene un módulo de minería de texto el cual realiza la extracción de frases clave y las adiciona como metadatos.¹³

Otro ejemplo sobre la aplicación de minería de texto es el software que utiliza métodos de agrupamiento o *clustering* para la organización de documentos en la biblioteca digital o en colecciones individuales, el cual agrupa los documentos en diferentes grupos temáticos basados en su contenido¹⁴; el resultado del agrupamiento se puede visualizar con el empleo del software libViewer¹⁵ como se muestra en la figura 2.

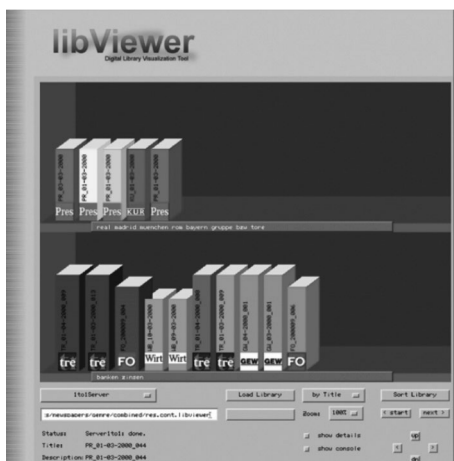


Figura 2. Visualización de la biblioteca digital de la sección de economía y deportes.

Etapas de la minería de texto

La minería de texto, como proceso, establece una serie de pasos necesarios para el procesamiento de texto y su consecuente extracción de datos/conocimiento;¹⁶ dependiendo del objetivo de la minería de texto los pa-

sos a seguir varían. A continuación se describen de manera general los más relevantes.

El primer paso es determinar el propósito de estudio de la minería de texto; como ejemplo, en el caso de textos de biología se puede identificar y etiquetar entidades biológicas, extracción y normalización de sinónimos, homónimos y abreviaturas, identificación de entidades biológicas, generación de hipótesis, etcétera.

El segundo paso es recolectar, identificar, recoger y validar información; en esta fase se realiza la recuperación de información (IR) en la cual se buscan e identifican las fuentes más relevantes para el objetivo de estudio de la minería de texto. Luego, se deben recopilar los documentos detectados en el mejor formato, se seleccionan, se evalúa su relevancia y se realizan las anotaciones necesarias.¹⁷ En esta etapa se cuenta con el conjunto de documentos necesarios para la realización de la minería de texto.

En el tercer paso se realiza el procesamiento de texto con la finalidad de eliminar información que no ayuda al propósito de la minería de texto, realizándose algunas de las siguientes acciones: análisis léxico, tratamiento y separación de palabras vacías (artículos, preposiciones, conjunciones), tratamiento de términos flexionados (términos relacionados morfológicamente, variaciones de género, número o tiempo verbal), tratamiento de palabras compuestas, normalización de palabras, obtención de las raíces de las palabras y etiquetado de palabras, además de corregir algunos problemas que presenten los documentos como: los problemas de formato, polisemia, homonimia, sinonimia. Esta fase exploratoria se basa principalmente en lingüística computacional (análisis morfológico y sintáctico), además de algoritmos informáticos.¹⁸ El objetivo de esta fase es facilitar la selección de características deseadas para identificar palabras clave,

¹³ Text mining in a digital library, *op. cit.*

¹⁴ GUPTA, Vishal, Lehal, Gurpreet S. A Survey of Text Mining Techniques and Applications.

¹⁵ RAUBER, Andreas, MERKL, Dieter. Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres.

¹⁶ SUKANYA, M., BIRUNTHA, S., *op. cit.*

¹⁷ *Extraction des informations et des connaissances* [en línea]. <<http://touriaelouahabi.wordpress.com/text-mining/principes-et-concepts-text-mining/>>

¹⁸ ABBOTT, Dean. *Introduction to Text Mining* [en línea]: *Virtual Data Intensive Summer School*, July 10, 2013. <<http://www.vscse.org/summerschool/2013/Abbott.pdf>>

identificación de entidades biológicas, individuos, organizaciones, lugares, oraciones, conceptos, etcétera.

En el cuarto paso se realiza la extracción y análisis de clases, relaciones, asociaciones o secuencias, con el fin de encontrar evidencias de conceptos y de estructuras existentes. En esta etapa los documentos se pueden representar a través del modelo del espacio vectorial,¹⁹ en donde cada documento es modelado como un vector de dimensión n y es representado de la siguiente manera

$D_i = (d_{i1}, d_{i2}, \dots, d_{in})$ donde cada d_{ij} representa el número de repeticiones de la palabra en el documento, también los datos obtenidos en esta etapa son representados en alguna estructura informática que facilita su análisis; las estructuras representan las relaciones entre las entidades de un mismo tipo de datos, palabras o conceptos clave, documento-términos, términos-autores, interacción de proteínas, etcétera.

En el último paso se presentan los resultados a través de resúmenes, marcados de texto, relaciones,

taxonomías y visualización para su interpretación. En esta etapa también se puede almacenar la información procesada en bases de datos para su recuperación posterior. La figura 3 muestra las etapas del proceso de minería de texto en el área biomédica para el reconocimiento de entidades.

Existe una diferencia entre recuperación de información y minería de texto. La recuperación de información es el primer paso en la minería de texto, es decir, una vez que ya se localizó un conjunto de documentos éstos sirven de base en los procesos de la minería de texto para extraer información e identificar patrones, lo que lleva al descubrimiento de nueva información con el objetivo de encontrar información que antes sólo pudo haber sido descubierta por la lectura de un gran número de documentos.

Algunos de los métodos utilizados en esta área emergente para el procesamiento de documentos textuales son: métodos de aprendizaje de máquina, procesamiento del lenguaje natural, redes neuronales artificiales,

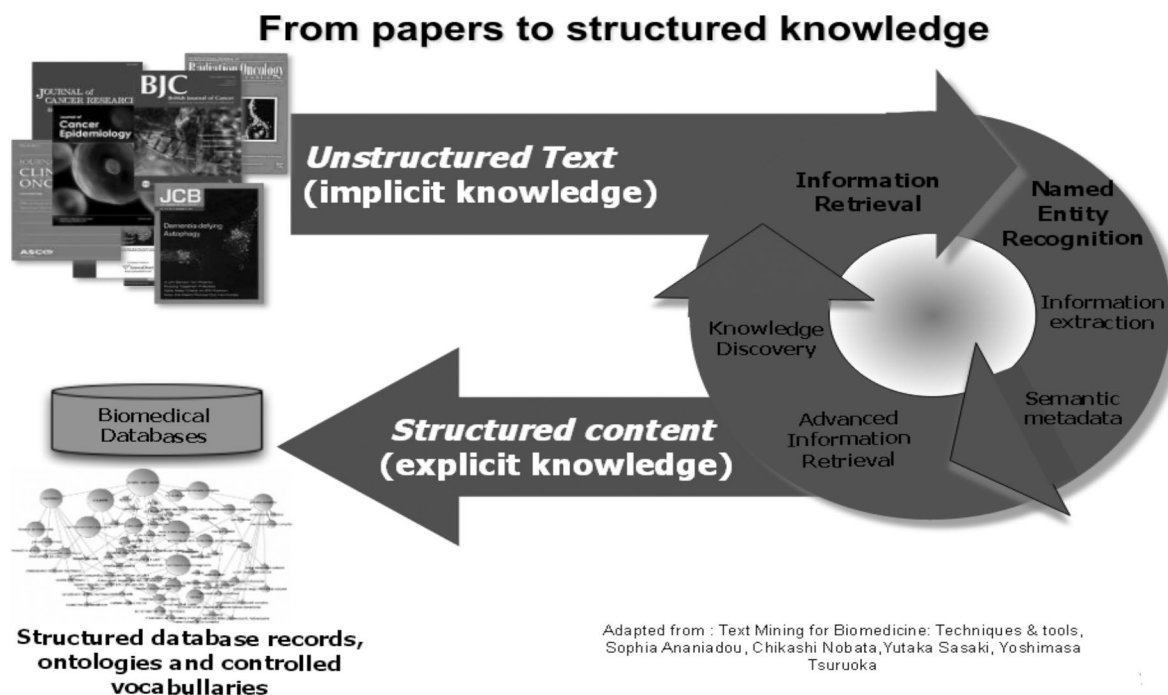


Figura 3: Etapas del proceso de minería de texto.^{20,21}

¹⁹ SALTON, Gerard. *Automatic text processing: The transformation, analysis, and retrieval of information by computer.*

²⁰ *Value and benefits of text mining* [en línea]. Jisc. <<http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>>

²¹ KRALLINGER, Martin. *Current trends in biomedical text mining: chemicals, drugs and mutations* [video]

extracción de la información, métodos matemáticos, modelo Bayesiano, modelos probabilísticos, modelo del espacio vectorial, Indexación Latente Semántica (LSI), entre otros. Algunos de los algoritmos de minería de texto empleados para clasificación, categorización y agrupación de información son los siguientes: cuantización vectorial, K-medias y Redes Neuronales Artificiales (RNA). Dentro de este último grupo se encuentran las RNA-Kohonen, Self-Organizing Map (SOM), Learning Vector Quantization (LVQ), entre otras.²² Todos utilizan un método competitivo, pero unos son no supervisados y otros son supervisados.

los digitales, por ejemplo en biomedicina, donde los volúmenes de información sobre temas específicos hacen que sea imposible leer todos los artículos por cualquier investigador y mucho menos estudiar los artículos relacionados. La figura 4 muestra algunas de las relaciones que se pueden establecer en el área de la biología.

Un ejemplo particular de la aplicación de los sistemas de minería de texto en el análisis de un documento se muestra en la figura 5, en donde el texto sombreado representa los posibles términos representativos del documento (los términos pueden estar formados por más de una palabra).

Aplicaciones de la minería de texto

Las técnicas de minería de texto se pueden utilizar para procesar diferentes tipos de documentos textua-

Otra de las aplicaciones es en el área médica, donde Swanson analizó artículos de la base MEDLINE para encontrar la relación existente entre la migraña y la deficiencia del magnesio; estos datos fueron validados

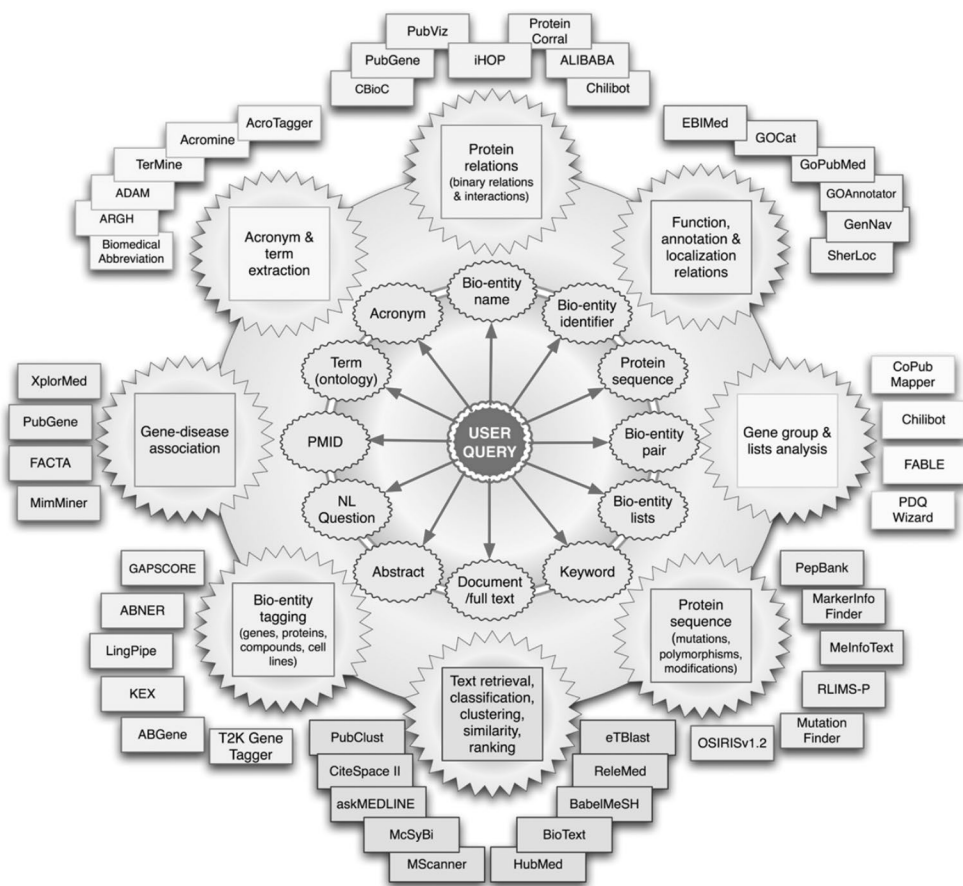


Figura 4: Minería de texto en biología.²³

²² HAYKIN, Simon. *Neuronal Networks: A Comprehensive Foundation*.

²³ KRALLINGER, Martin, *op. cit.*

AcroMine is based on a novel approach for recognising acronym definitions in a text collection. Applied to the whole MEDLINE with terminological variation, an integral part of the linguistic ability to realise a concept in different ways. This is also an obstacle for information retrieval. AcroMine finds expanded forms of acronyms from a database created from the whole of Medline and disambiguates them. AcroMine has also been incorporated into the TerMine user interface. TerMine, moreover, crucially supports Semantic Web activities and has been used as an aid to ontology construction and controlled vocabularies.

Cheshire / TerMine integrates information retrieval (provided by Cheshire) and TerMine to offer users a search facility based on which they are familiar, and which moreover retrieves documents according to the importance of the terminology they choose. For example, this service allows the user to search for documents related to a subject area, to find associated terms related with it and to select the most appropriate documents to view. This service identifies important terms (that the user may not have known) combining the best of search and browse models of information access.

MEDIE provides real-time semantic information retrieval based on the retrieval of relational concepts from huge texts. It is an advanced search engine which uses semantic retrieval technologies to identify sentences containing biomedical correlations for queries from Medline abstracts. The service runs on the whole of Medline and is based on semantically annotated texts using deep parsing and named entity recognition. Sentences are annotated in advance with semantic structures and stored in a structured database. User requests are converted on the fly into patterns of these semantic annotations, and texts are retrieved by matching these patterns with the pre-computed semantic annotations.

InfoPubMed extracts and visualises protein-protein interactions. Info-PubMed is an efficient PubMed search tool, helping users obtain information about biomedical entities such as genes, proteins, and the interactions between them.

Figura 5: Extracción de términos por métodos de minería de texto.²⁴

posteriormente de manera experimental.^{25,26} Además, la minería de texto se ha utilizado para el descubrimiento de fármacos, la toxicología predictiva, la inteligencia competitiva, la búsqueda de patentes, etcétera.

Max Haeussler, un investigador biólogo de la Universidad de California, utilizó documentos científicos de la base de datos de Elsevier para extraer datos sobre ADN en cerca de 3 millones de artículos con el fin de realizar el mapa del genoma humano.

En el área biomédica la minería de texto es aplicada en las siguientes tareas: reconocimiento de compuestos químicos y medicamentos, aplicación en la extracción de drogas y sus efectos, para determinar reacciones enzimáticas, en la clasificación de las mutaciones y las proteínas del cáncer, en la detección de especies u organismos.²⁷

Otros proyectos, como las investigaciones sobre el cerebro humano desarrolladas por el consorcio europeo integrado por investigadores que trabajan en el proyecto cerebro humano, están utilizando la interfaz de Elsevier para realizar los trabajos de minería de texto.

Un ejemplo más del uso de los artículos de Elsevier en la aplicación de la minería de texto es el proyecto del estudio de la fisiología de la neurona.²⁸

También se puede emplear para: la inteligencia del gobierno y las agencias de seguridad tratando de reconstruir las advertencias terroristas y otras amenazas de seguridad, el monitoreo en redes sociales, la inteligencia en los negocios, el análisis sentimental; este último se refiere a la exploración de texto para identificar y extraer actitudes del escritor de un texto, identificar el estado emocional del autor o la intención emocional de comunicación. La tarea del análisis sentimental es clasificar un documento o una oración para identificar características positivas, negativas o neutras y también es utilizado para determinar estados emocionales como el enojo, la tristeza y la felicidad. El análisis sentimental puede ser aplicado en la Web, las redes sociales y los blogs.²⁹

Debido a la importancia que ha adquirido la minería de texto algunas instituciones académicas hacen un esfuerzo para que esta área se desarrolle, como es el caso del Centro Nacional de Minería de Texto (NACTeM, por sus siglas en inglés), el cual es operado por la Universidad de Manchester en estrecha colaboración con la Universidad

²⁴ Ejemplo generado con el software TerMine: sitio <http://www.nactem.ac.uk/software/termine/#form>

²⁵ SWANSON, Don R., *op. cit.*

²⁶ SMALHAISER, Neil, Swanson, D. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease.

²⁷ KRALLINGER, Martin, *op. cit.*

²⁸ NOORDEN, Richard Van. Elsevier opens its papers to text-mining.

²⁹ *Sentiment analysis* [en línea]. Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=625073725>

de Tokio, apoyado con fondos públicos. Este centro tiene como objetivo la promoción de la investigación de las tecnologías de minería de texto y su uso; ofrece herramientas de software para algunas tareas de minería de texto, seminarios, conferencias, tutoriales, demostraciones y publicaciones.³⁰ En el sector privado también se están realizando investigaciones, muestra de ello son las compañías de fármacos que utilizan la minería de texto para reducir los costos de sus investigaciones.³¹

Además, empresas como Elsevier han puesto al servicio de los científicos más de 11 millones de artículos de investigación disponibles en línea, para ser utilizados en tareas de minería de texto, para extraer hechos y relaciones de éstos. En tiempos pasados la empresa no permitía hacer una recolección de los artículos publicados en su sitio, impidiendo así la aplicación de la minería de texto, muestra de ello es el caso de Max Haeussier, de la Universidad de California, que invirtió más de tres años para conseguir el permiso para extraer datos sobre el ADN. Es por eso que el pasado 26 de enero de 2014, en la conferencia de la American Library Association en Philadelphia, Pennsylvania, Elsevier anunció que instituciones académicas y de investigación pueden utilizar la interfaz de Elsevier (API) para descargar documentos en formato XML de forma masiva, con un máximo de 10 000 artículos por semana, y pueden ser utilizados libremente para la minería de texto siempre y

cuando las instituciones académicas y de investigación establezcan un acuerdo legal. Los acuerdos incluyen que los investigadores pueden publicar sus productos de minería de texto mientras no sea para uso comercial y pueden incluir máximo 200 caracteres del texto original y deben hacer referencia al contenido original.³²

Software de minería de texto

En la actualidad existe una amplia gama de software para realizar minería de texto, tanto de tipo comercial como de acceso libre; dentro del tipo de acceso libre se mencionan los siguientes: Gate, con capacidad de resolver problemas de procesamiento de texto; Carrot, utilizado para organizar de forma automatizada documentos en categorías temáticas; RapidMiner, utilizado para el aprendizaje de máquina, minería de datos y texto. De tipo comercial son: SAS Text Miner, para extraer información de datos no estructurados; Clearforest, utilizado para el análisis de texto y procesamiento de lenguaje natural; Autonomy, realiza minería de texto, categorización y clasificación de documentos; SPSS, dentro de sus funciones realiza minería de texto y procesamiento de lenguajes natural; Lexalytics, utilizado para procesamiento de texto, procesamiento de lenguaje natural y extracción de entidades. El URL de los desarrolladores de minería de texto se muestra en la tabla 1.

Tabla 1: Software comercial para minería de texto³³

Vendedor	Sitio Web	Producto
Autonomy	http://www.autonomy.com/products/	IDOL Server, Retina
Clearforest	http://www.clearforest.com/solutions.html	ClearForest Text Analysis Suite
SAS	http://www.sas.com/en_us/software/analytics/text-miner.html	SAS Text Miner
IBM Text Analytics	http://www-ibm.com/software/ebusiness/jstart/textanalytics	IBM
SPSS	http://www-01.ibm.com/software/analytics/spss/	LexiQuest, Clementine
Lexalytics	http://www.lexalytics.com/web-demo	Text Analytics Software
Gate	http://gate.ac.uk/download/	Gate
Carrot	http://project.carrot2.org/index.html	Carrot Search
RapidMiner	http://sourceforge.net/projects/rapidminer/files/2. Extensions/Text Processing	RapidMiner

³⁰ *The National Centre for Text Mining* [en línea]. <<http://www.nactem.ac.uk/>>

³¹ BELSKY, Gary. Why Text Mining May Be The Next Big Thing [en línea]. *Technology & Media*, 20 de Marzo de 2012. <<http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/>>

³² NOORDEN, Richard Van, *op. cit.*

³³ Tapping the Power of Text Mining. Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang.

La diferencia en esta gama de software es que algunos realizan extracción de información, seguimiento de temas, resúmenes, categorización, agrupamiento, respuestas a preguntas, extracción de términos.³⁴ Por otra parte, algunos pueden ser utilizados a través de la Web y otros requieren instalación a nivel local. También existen desarrollos de sistemas de cómputo en las universidades cuya finalidad es poder programar los algoritmos de minería de texto, de acuerdo al tipo de información y a la necesidad particular de procesamiento de información.

Como podemos ver, la aplicación de la minería de texto es una herramienta de importancia relevante en la actualidad. Independientemente de la disciplina en la cual se aplique, nos permite localizar y descubrir información oculta en grandes volúmenes de información de manera automatizada, permitiendo analizar documentos con la finalidad de obtener nueva información para generar nuevo conocimiento o generar hipótesis de investigación.

Comentarios finales

El desarrollo tecnológico alcanzado en los últimos años ha traído como consecuencia una producción acelerada de la información en formato digital, lo que

genera la necesidad de resolver problemas y retos para el procesamiento de la misma. Por tal razón, es preciso contar con los métodos y las tecnologías necesarios para su procesamiento, organización y clasificación, entre otras tareas.

La minería de texto ayuda a explorar y explotar los contenidos de los documentos textuales digitales, que están en continuo incremento, con la meta de obtener información relevante, sirviendo de base en el desarrollo de investigaciones académicas y de negocios que debido a su potencial puede ser aplicada en todas las disciplinas; el área en donde más aplicaciones y desarrollo tienen es en las ciencias biomédicas y campos relacionados, sin embargo, la minería de texto en otros campos del conocimiento no ha tenido gran aplicación. Las tecnologías de la minería de texto pueden ser utilizadas también en las bibliotecas en el procesamiento y análisis de contenidos, con el objetivo de obtener relaciones entre temas y autores de los documentos académicos. En los ambientes académicos existen algunos obstáculos que deben de superarse, como los derechos de copia y las restricciones de acceso, en beneficio del desarrollo de la minería de texto. ❧

Obras consultadas

ABBOTT, Dean. *Introduction to Text Mining* [en línea]. *Virtual Data Intensive Summer School*, July 10, 2013. <<http://www.vscse.org/summerschool/2013/Abbott.pdf>> [Consulta: 17 junio 2014].

BEATH, Cynthia, BECERRA-FERNÁNDEZ, Irma, ROSS, Jean, SHORT, James. El valor de la explosión de la información [en línea]. *IQ Intelligence quarterly: diario de análisis avanzado*, 2013, p. 8-11. <http://www.sas.com/offices/europe/spain/IQ/Q113/IQBigAnalytics_GrandesOportunidades.pdf>

BELSKY, Gary. Why Text Mining May Be The Next Big Thing [en línea]. *Technology & Media*, 20 de Marzo de 2012. <<http://business.time.com/2012/03/20/why-text-mining-may-be-the-next-big-thing/>>

Digital Universe [en línea]. EMC2. <<http://www.emc.com/leadership/programs/digital-universe.htm>> [Consulta: 22 octubre 2013].

Extraction des informations et des connaissances [en línea]. <<http://tourielouahabi.wordpress.com/text-mining/principes-et-concepts-text-mining/>> [Consulta: 17 junio 2014].

FUNG, Benjamin C.M., Wang, Ke, Ester, Martin. B. *Hierarchical Document Clustering Using Frequent Itemsets* [en línea]. <<http://epubs.siam.org/doi/pdf/10.1137/1.9781611972733.6>> [Consulta: 14 julio 2010].

Fuzzy cluster analysis. Frank Höppner y otros. New York: J. Wiley, c1999. 289 p.

³⁴ *Idem*.

- GUPTA, Vishal, LEHAL, Gurpreet S. A Survey of Text Mining Techniques and Applications. *Journal of emerging technologies in Web intelligence*, August 2009, vol. 1, no. 1, p. 60-76.
- HAYKIN, Simon. *Neuronal Networks: A Comprehensive Foundation*. New York: Macmillan, 1994.
- Journal Selection for MEDLINE Indexing at NLM* [en línea]. U.S. National Library of Medicine. <http://www.nlm.nih.gov/pubs/factsheets/j_sel_faq.html#a15> [Consulta: 22 octubre 2013].
- KRALLINGER, Martín. *Current trends in biomedical text mining: chemicals, drugs and mutations* [video]. Madrid, España: S. N. Centre, 2013. Jornadas Mavir, (19 de noviembre de 2013).
- MARÍN, María José, RICO, Eva, JULI, Pascual. *Data mining en relación a la documentación periodística* [en línea]. <<http://personales.upv.es/ccarrasc/doc/2003-2004/DMPeriodistico/TREBALLSRP.htm>> [Consulta: 2011].
- The National Centre for Text Mining* [en línea]. <<http://www.nactem.ac.uk/>> [Consulta: 20 marzo 2014].
- Neural Networks Design*. Martin T. Hagan, Howard B. Demuth, Mark H. Beale, Orlando de Jesús. E.U.A.: PWS Publishing, 1996.
- NOORDEN, Richard Van. Elsevier opens its papers to text-mining. *Nature*, 3 febrero 2014, vol. 506, no. 7486.
- RAUBER, Andreas, MERKL, Dieter. Text Mining in the SOMLib Digital Library System: The Representation of Topics and Genres. *Applied Intelligence*, 2003, vol. 18, p. 271-293.
- SALTON, Gerard. *Automatic text processing: The transformation, analysis, and retrieval of information by computer*. Reading, Massachusetts: Addison Wesley, c1989. 530 p.
- SÁNCHEZ, D., Martín-Bautista, M. *Un enfoque deductivo para la minería de texto* [en línea]. <<http://www.softcomputing.es/estylf08/es/2006-XIII%20Congreso/articulos/40.pdf>> [06 febrero 2014].
- Sentiment analysis* [en línea]. Wikipedia, The Free Encyclopedia. <http://en.wikipedia.org/w/index.php?title=Sentiment_analysis&oldid=625073725> [Consulta: 18 septiembre, 2014].
- SMALHAISER, Neil, SWANSON, D. Assessing a gap in the biomedical literature: magnesium deficiency and neurologic disease. *Neuroscience research communications*, 1994, vol. 15, p. 1-9.
- SUKANYA, M., BIRUNTHA, S. Techniques on Text Mining. En: *IEEE International Conference on Advanced Communication Control and Computing Technologies (ICACCT)*, 23-25 Aug., 2012. Ramanathapuram: IEEE, 2012.
- Survey of Text Mining II: Clustering, Classification and Retrieval*. Michael W. Berry, Malu Castellanos, editors. London: Springer, 2008. 240 p.
- SWANSON, Don R. Complementary structures in disjoint science literatures. En: *Proceeding SIGIR '91, Proceedings of the 14th Annual International ACM/SIGIR Conference on Research and development in information retrieval*. New York: ACM, c1991, p. 280-289.
- Tapping the Power of Text Mining. Weiguo Fan, Linda Wallace, Stephanie Rich, Zhongju Zhang. *Communication of the ACM*, September 2006, vol. 49, no. 9, p. 76-82.
- Text mining: A new frontier for lossless compression. Ian H. Witten, Zane Bray, Malika Mahoui, Bill Teahan. En: *Data Compression Conference, 1999. Proceedings. DCC '99*. Snowbird: IEEE, 1999, p. 198-207.
- Text mining in a digital library. Ian H. Witten, Katherine J. Don, Michael Dewsnip, Valentin Tablan. *International Journal on Digital Libraries*, 2003, vol. 4, no. 1, p. 56-59.
- Value and benefits of text mining* [en línea]. Jisc. <<http://www.jisc.ac.uk/reports/value-and-benefits-of-text-mining>>
- YANLIANG QI, Zhang, YANG, Song, Min. Text Mining for Bioinformatics: State of the Art Review. En: *Conference on Computer Science and Information Technology, 2009, ICCSIT 2009*, 2nd IEEE International Conference on Computer Science and Information Technology, 8-11 Aug., 2009. Beijing: IEEE, 2009.