Text Mining and BI

Abstract

Los recientes avances en lingüística computacional, así como la tecnología de la información en general, permiten que la inserción de datos no estructurados en una infraestructura de inteligencia de negocios resulte viable y eficaz. Dichos avances, integrados, son considerados como lo que se denominaría "minería de texto", que consiste en el ***análisis de texto en su lenguaje natural con el fin de extraer términos clave, entidades y relaciones entre esos términos y entidades****. Estos elementos extraídos son utilizados para diversos fines, entre ellos:*

- *Categorización y clasificación de documentos*
- *Generación de resúmenes de textos*
- *Proporción de datos que permiten la visualización de herramientas que sirven para navegar importantes bases de datos de texto y*
- *Mapeo de relaciones multifase.*

El "Text Mining"

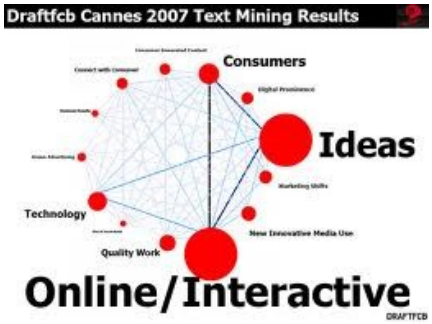
Dado que la demanda de técnicos de inteligencia de negocios más efectivos va en aumento, los profesionales de BI están viéndose en la necesidad de encontrar bases informáticas más amplias a fin de incluir texto no estructurado.

Las técnicas de minería de texto son, por lo tanto, esenciales para explotar las fuentes de de información.

Las tres técnicas fundamentales sobre minería de texto en inteligencia de negocios son:

1. la extracción de términos,
2. la extracción de información y
3. el análisis de enlaces.

Tradicionalmente, la inteligencia de negocios se ha centrado en el análisis de los datos obtenidos de los sistemas de procesamiento de transacciones, como la planificación de recursos empresariales (ERP), la gestión de relaciones con clientes (CRM), la automatización de fuerza de ventas (SFA), el procesamiento de quejas y reclamos, y otras fuentes de datos estructurados.



Las bases de datos relacionales y las orientadas a objetos, son utilizadas comúnmente para implementar estos modelos.

En los datos no estructurados, como su nombre lo indica, no existe un esquema formal para los mismos.

El texto, el audio y el video en formato libre son las formas más comunes que existen para este tipo de datos no estructurados.

Los datos no estructurados son en realidad los que más abundan en la mayoría de las organizaciones, pero hasta el día de hoy no han sido aprovechados adecuadamente como fuente de inteligencia de negocios.

Ejemplo:

Consideremos la siguiente posible toma de decisión. Un Director de Ventas de una compañía de telecomunicaciones planifica una nueva promoción de servicios inalámbricos y necesita obtener información sobre:

- Desempeño de promociones anteriores.
- Tendencias generales en las ventas de la compañía.
- Nuevos factores de mercado que influyen en las decisiones de compra.
- Capacidad dentro de la propia infraestructura de la empresa, y
- comportamiento histórico de los competidores y planificación para promociones similares.

Un Datamart tradicional de ventas con información sobre transacciones responderá fácilmente a aquellas dudas sobre promociones anteriores y tendencias generales.

Un Datamart de operaciones abordaría de manera similar la capacidad y las dudas de infraestructura. Éstos, sin embargo, son indicadores históricos y, aunque necesarios para el proceso de toma de decisiones, no son suficientes.

El mercado de las telecomunicaciones suele ser muy frágil debido a las tecnologías, la inclusión de intermediarios especializados, por ejemplo de telefonía celular prepaga, y variaciones en la demografía de los clientes (por ejemplo, el aumento de los clientes adolescentes). Algunos de estos factores se reflejan en los sistemas tradicionales de inteligencia de negocios, pero sólo después de cruzar un umbral medible. Para detectar y evaluar el impacto de estos factores, se requiere con antelación el análisis de fuentes basadas en texto, tales como los informes internos sobre futuros e información sobre marketing, noticias del sector, informes gubernamentales y otras fuentes internas y externas.

Volviendo a nuestras tres técnicas de data mining, podemos comenzar por la **extracción de términos**, la técnica más básica, que identifica los términos clave y entidades lógicas, como los nombres de las organizaciones, lugares, fechas y valores financieros entre otros. La Extracción de términos es el formato más básico de minería de texto. De igual manera que todas las técnicas de minería de texto, éste detecta información a partir de datos no estructurados dentro de un formato estructurado. La estructura de datos más simple en la minería de texto es el vector de características,

una lista de las palabras ponderadas que aparecen en un texto. Proporciona una descripción representativa, o la firma para el texto.

La **extracción de información**, la segunda técnica mencionada, se basa en los términos extraídos del texto para identificar las relaciones básicas, tales como las funciones de las distintas empresas en una fusión, o la promoción de la reacción química de una enzima. Entonces, el siguiente nivel de complejidad en la minería de textos es la extracción de información. A diferencia de la extracción de términos que se centra en las condiciones, la extracción de información se centra en un conjunto de hechos que constituyen un evento, episodio, o estado.

Por último, el **análisis relacional**, combina múltiples vínculos para formar modelos de varios pasos de procesos complejos, tales como las vías metabólicas. El análisis de enlaces es un conjunto de técnicas que permite tener una idea de las relaciones entre varias entidades con múltiples conexiones, pasos, o enlaces.

En conjunto, estas tres técnicas proporcionan el fundamento para la integración de inteligencia de negocios basada en texto dentro de los sistemas de BI existentes.

Beneficios de la Minería de Datos de Texto

Es importante diferenciar entre Minería de Datos de Texto y acceso a la información.

El objetivo de acceso a la información es ayudar a los usuarios a encontrar los documentos que satisfagan sus necesidades de información[1]. Sin embargo, la minería de texto se centra en cómo usar un bloque de información textual, como una amplia base de conocimientos a partir de la cual se puede extraer nueva información nunca antes conocida[2].

Además de proporcionar herramientas para ayudar en el proceso de acceso a la información estándar, la minería de texto puede contribuir a proporcionar sistemas suplementados con herramientas para el análisis exploratorio de los datos.

Otra forma de entender la minería de texto sería abordándola como un proceso de análisis exploratorio de datos[3] que conduce al descubrimiento de información hasta el momento desconocida, o utilizada para responder a preguntas para las cuales no se conoce actualmente la respuesta. Se podría citar como ejemplo, estudios hechos sobre texto de literatura médica y el impacto social de la minería de texto.

La extracción de características, pertenece a la minería de texto dentro de un documento que trata de encontrar vocabulario significativo e importante dentro de un documento de texto en lenguaje natural. Esto implica el uso de técnicas que incluyen la búsqueda de patrones y heurística que se centran en el léxico y parcialmente en la información del lenguaje. Por otro lado, podría encontrarse el Hipertexto e Hipermedia de minería de datos, así como la Minería de Datos Visual y la Minería de Datos Multimedia.

Conclusión

La Inteligencia de negocios demanda una amplia gama de fuentes de datos, que no son solamente numéricos.

Las tres técnicas básicas, Términos y extracción de características; Extracción de información; y Análisis Vinculados proporcionan las bases de las técnicas de inteligencia de negocios relacionadas con el texto. Estas técnicas detectan elementos clave de texto de forma libre en formatos estructurados los cuales se prestan a sí mismos a las técnicas de análisis.

La Extracción de términos, con requisitos de procesamiento lingüístico relativamente simples, aprovecha el análisis estadístico para medir la relevancia de los términos que a su vez representan conceptos en una colección. La Extracción de información identifica ambas entidades relevantes y sus relaciones. El análisis de enlaces es uno de los métodos que se utiliza para estudiar la relación entre los eventos derivados de las técnicas de extracción de información.

A diferencia de las técnicas tradicionales de inteligencia de negocios, éstas son pasibles de sufrir fallas, incluso cuando se programan correctamente.

El estado de la técnica en el procesamiento del lenguaje no es suficiente para analizar con precisión la variedad de temas y estilos que se encuentran en el entorno empresarial. Incluso con las limitaciones actuales, las técnicas de minería de texto pueden proporcionar información valiosa de la que no se tiene acceso a través de los datos estructurados existentes.

Ing. Sergio D. Salimbeni

[1] Baeza-Yates & Ribeiro-Neto, 1999

[2] Craven et al.,1998

[3] Tukey, 1997

[4] Sullivan Dan, The Ballston Group.