PONTIFICIA UNIVERSIDAD CATOLICA DE VALPARAISO
FACULTAD DE INGENIERIA
ESCUELA DE INGENIERIA INFORMATICA

UN FRAMEWORK PARA LA DETECCIÓN DE FRAUDE POR SUBVALORACIÓN DE PRODUCTOS MEDIANTE TÉCNICAS DE MINERÍA DE TEXTO Y DATOS

LUIS ALBERTO SUÁREZ PIZARRO

TESIS DE GRADO
MAGÍSTER EN INGENIERÍA INFORMÁTICA

(Septiembre 2011)

PONTIFICIA UNIVERSIDAD CATOLICA DE VALPARAISO
FACULTAD DE INGENIERIA
ESCUELA DE INGENIERIA INFORMATICA

UN FRAMEWORK PARA LA DETECCIÓN DE FRAUDE POR SUBVALORACIÓN DE PRODUCTOS MEDIANTE TÉCNICAS DE MINERÍA DE TEXTO Y DATOS

LUIS ALBERTO SUÁREZ PIZARRO

Profesor Guía: **José Luis Martí Lara**

Programa: **Magíster en Ingeniería Informática**

(Septiembre 2011)

RESUMEN

El presente trabajo pretende demostrar que por medio de la utilización conjunta de técnicas de minería de texto y datos es posible enriquecer un proceso de clasificación orientado a la detección de fraude, específicamente, aquel asociado a la subvaloración de productos importados. El estudio se centra en la definición de un *framework* que utiliza técnicas de minería de texto para la clasificación de productos, a partir de la descripción de éstos, la cual se encuentra como texto no estructurado, donde se realiza una descripción en lenguaje natural para definir las características de los productos. Posteriormente, los productos identificados a partir de la clasificación anterior, pasan por un proceso de minería de datos con el fin de generar agrupamientos en base a atributos de interés, para poder identificar elementos con alta probabilidad de fraude utilizando técnicas de detección de outliers.

Términos claves: Minería de Datos, Minería de Texto, Agrupamiento, Detección de Fraude, Outliers.

ABSTRACT

This thesis tries to prove that, through the use of techniques of text and data mining, it is possible to improve a classification process oriented to detect fraud, specifically the one associated to the undervaluation of imported products. This research focuses on the definition of a framework that could use text mining to classify products from their own description. Descriptions are in a state of unstructured text, and in order to define the product features, these are made in natural language. Subsequently, the identified products from these descriptions pass through a data mining process in order to generate clusters based on attributes of interest. The use of techniques of *outlier* detection allows to identify elements with high possibilities of fraud.

Key words: Data Mining, Text Mining, Clustering, Fraud detection, Outliers.

ÍNDICE

Índice de figura

Índice de tablas

1.- INTRODUCCIÓN

En el Servicio Nacional de Aduanas (SNA) se procesan miles de transacciones diarias asociadas a tramitaciones de comercio exterior que se realizan por medio del envío de documentos electrónicos, en una modalidad B2G entre el SNA y los agentes de Aduanas. Estas son registradas y posteriormente almacenadas en bases de datos históricas con el fin de realizar análisis posteriores.

El masivo flujo de transacciones es producto de la madurez tecnológica de la organización, la que actualmente tienen gran parte de los procesos de importaciones y exportaciones automatizados, gracias a la utilización de documentos electrónicos. Esto ha producido que los tiempos de respuesta asociados al procesamiento de las tramitaciones disminuya notablemente, y al mismo tiempo, se puedan incorporar mecanismos de control en base a la utilización de filtros, los cuales se activan cuando ciertos campos de las transacciones de importaciones, que contemplan productos controlados, sobrepasan los límites establecidos por el SNA.

Lamentablemente, como los procesos de fiscalización se realizan en gran parte de forma manual, a medida que el volumen de importaciones y exportaciones aumenta, se hace más difícil realizar estas tareas de una manera eficiente. Ya que las importaciones aumentan cada año, a diferencia del los recursos humanos que se mantienen prácticamente constantes y en una clara desproporción frente a la actual demanda.

Como no es posible fiscalizar la totalidad de lo que entra y sale del país, un enfoque aceptable es disponer de herramientas que permitan detectar aquellas transacciones que por sus características, pudieran incorporar algún tipo de fraude. De esta forma, se focalizan los escasos recursos en las situaciones en que se justifique incurrir en los costos asociados a un proceso de fiscalización, y al mismo tiempo, se trabaja con transacciones previamente seleccionadas por los modelos de clasificación de fraude, las que cumplen con presentar una alta probabilidad de resultar positivas. Cabe destacar, que si estas herramientas poseen un elevado nivel de efectividad en su proceso de detección, se transforman en elementos valiosos para las tareas de inteligencia aduanera y por ende, para el proceso de fiscalización.

Por las razones expuestas, este trabajo se centra en la confección de un *framework* el cual cubre aspectos metodológicos, teóricos y prácticos, los cuales son incorporados en una solución donde además se incorporan tanto productos de software propietarios como *open sources*, los cuales logran trabajar en conjunto por medio del desarrollo de diversos componentes, los cuales se encargan de realizar gran parte de las tareas de procesamiento y al mismo tiempo permiten la cohesión de los diversos componentes. El producto se focaliza en la detección de fraude por subvaloración de productos desde un punto de vista práctico, con la finalidad de demostrar que con un enfoque apropiado se pueden enfrentar una gran cantidad de casos de fraudes, sin caer en el *overhead* de incorporar complejos mecanismos para tratar de identificar comportamientos fraudulentos.

El diseño del *framework* se basa en la incorporación de técnicas obtenidas desde distintas áreas tales como: Procesamiento de lenguaje Natural, Recuperación de Información, Extracción de Información, minería de texto y de

datos. Un trato especial reciben las técnicas asociadas a minería de texto, la detección de anomalías como un mecanismo eficaz para la detección de fraude y aquellas herramientas que permiten la creación de grafos para ser utilizadas en minería gráfica.

En resumen, esta tesis pretende demostrar que los fraudes de subvaloración de productos en las importaciones pueden ser enfrentados por medio de un *framework* que contemple un planteamiento metodológico y al mismo tiempo incorpore la utilización de técnicas de minería de texto y datos, con el fin de clasificar aquellas importaciones que presenten una alta probabilidad de ser consideradas fraudulentas.

1.1. OBJETIVOS

1.1.1. Objetivo General

Este trabajo teórico-práctico tiene por objetivo principal, el estudio y propuesta de técnicas asociadas al área de la minería de texto y datos, con el fin potenciarlas e incorporarlas en la construcción de un *framework*, destinado a recuperar información de campos no estructurados para utilizarla en la generación de agrupaciones de productos con atributos homogéneos, y así facilitar la identificación de *outliers* que tengan alta probabilidad de estar asociados a comportamiento de fraude en la subvaloración de importaciones.

1.1.2. Objetivos Específicos

Para lograr el objetivo general, es necesario cumplir los siguientes objetivos específicos:

- Desarrollar un estudio del estado del arte en minería de texto y datos.
- Realizar la recolección de información, analizando las fuentes de datos disponibles y la calidad de los mismos.
- Eliminar parte del ruido presente en los datos recurriendo a técnicas de procesamiento de lenguaje natural y modelos de lenguaje.
- Utilizar técnicas de minería de texto para extraer una descripción estructurada de productos a partir de datos no estructurados.
- Utilizar técnicas de agrupamientos y representaciones gráficas para la detección de *outliers*, con el fin de utilizar estos resultados para identificar casos probables de fraude por subvaloración de productos.

1.2. ESTRUCTURA DE LA TESIS

El presente trabajo de tesis se estructura en seis capítulos. El primero corresponde a la introducción y planteamiento de objetivos. El segundo cubre los aspectos del negocio que son importantes para entender el problema de fondo y el contexto sobre el cual se va a trabajar, el que comprende la terminología técnica y la descripción del proceso motivo de estudio. En el capítulo tres, que junto al cuarto corresponden al núcleo de esta memoria, se realiza una revisión del estado del arte en los campos de minería de texto y de datos, conocimientos que son necesarios para posteriormente seleccionar las técnicas que serán utilizadas. En el capítulo cuarto, se realiza una descripción detallada de los

componentes que forman parte del *framework*, distinguiendo entre éstos a los que corresponden al campo de la minería de texto o de la minería de datos, y se detallan además las salidas a obtener. El capítulo quinto comprende las evaluaciones previas sobre las fuentes de datos y aquellas asociadas a los módulos que componen el *framework*. Finalmente en el capitulo seis y siete, se entregan las conclusiones, discusiones y el trabajo futuro.

2. DEFINICIÓN DEL PROBLEMA

En este capítulo se realiza una breve introducción sobre aquellos conceptos que son propios del negocio, los cuales tienen que ser mencionados previamente para poder comprender el contexto sobre el cual se realizará el estudio. Para facilitar la comprensión, se comienza por detallar los aspectos genéricos para posteriormente profundizar en los aspectos técnicos asociados al proceso de fiscalización. Todo esto, con el fin de aportar los elementos necesarios para definir un marco de trabajo que pueda ser claramente acotado al comienzo de este informe.

2.1. El SERVICIO NACIONAL DE ADUANAS (SNA)

Antes de profundizar en el problema de fondo, es necesario contar con una visión general de la organización sobre la cual se pretende aplicar los resultados que se obtengan de este trabajo. Por esta razón, la presente sección tiene por objetivo interiorizar al lector sobre las características del negocio, los problemas existentes y el proceso de negocio en el cual se enmarcará el trabajo realizado.

2.1.1. Las Funciones del Servicio

Cuando se hace referencia a la función de Aduanas dentro del país, la función más obvia que se viene a la mente es la que tiene relación con la presencia física en la frontera para la "vigilancia", es decir el control del movimiento o paso de las mercancías y pasajeros. Pero la función aduanera va más allá de esta actividad, ya que dentro de su cometido se encuentran implícitas las siguientes áreas de acción:

- **Misión fiscalizadora**: Tiene relación con la determinación y el control, en diferentes grados, de la aplicación y pago de los aranceles e impuestos que gravan el comercio exterior. Esta área se caracteriza por tener como cliente directo al Fisco, y como producto principal la recaudación fiscal por medio del cobro de aranceles.
- **Misión comercial**: Se basa en la aplicación de la ley para controlar los delitos contra la propiedad intelectual. También se preocupa de la implantación de medidas para el control del comercio según las regulaciones y normas técnicas existentes. Y realiza el control administrativo de aquellos agentes económicos facilitadores del comercio exterior tales como agentes de Aduana, almacenistas, *forwarders,* y en general todos aquellos que se encuentran bajo jurisdicción administrativa del servicio.
- **Protección social**: Está centrada en aplicar las medidas necesarias para controlar el ingreso y salidas de especies asociadas a narcotráfico, tráfico de armas, mercancías peligrosas, especies protegidas y objetos de patrimonio cultural, entre otras.

Se puede concluir que el carácter principal de las diversas responsabilidades que tiene una Aduana frente a su Gobierno, y en definitiva frente a la comunidad, es ser una entidad reguladora, la cual además de realizar la tarea de recaudación de impuestos se preocupa de llevar a cabo un servicio de control y protección.

2.1.2. Modelo de Gestión de Riesgo

En el año 1994 el Fondo Monetario Internacional (FMI) realizó un estudio sobre el estado de modernización del SNA, el cual concluyó que los sistemas de control, se caracterizaban por excesivas revisiones físicas y trámites junto a un escaso control efectivo [6]. En base a estos resultados se instó al Estado para modernizar el sistema de fiscalización el que evolucionaría desde un sistema aleatorio, principalmente físico, a uno basado en la evaluación de riesgo centrado en el análisis a posteriori de las mercancías, con el fin de facilitar el desarrollo del comercio.

La gestión de riesgo se puede definir como "el proceso de toma de decisiones en un ambiente de incertidumbre sobre una acción que va a suceder y sobre las consecuencias que existirán si esta acción ocurre". El modelo de gestión de riesgo tiene como objetivo, en el contexto del comercio exterior, determinar los peligros asociados con el propósito de decidir sobre las acciones de fiscalización que debe realizar el SNA.

El proceso de gestión de riesgo (representado en la Figura 1) es el utilizado actualmente para la identificación de riesgos en el área de fiscalización, el cual se basa en el estándar internacional AS/NZS/4360/99. Sin embargo, este proceso no está limitado exclusivamente a esta área operacional; por las características del mismo, éste se puede aplicar a cualquier contexto donde resulte de interés la identificación y tratamiento de potenciales peligros que necesitan ser manejados.

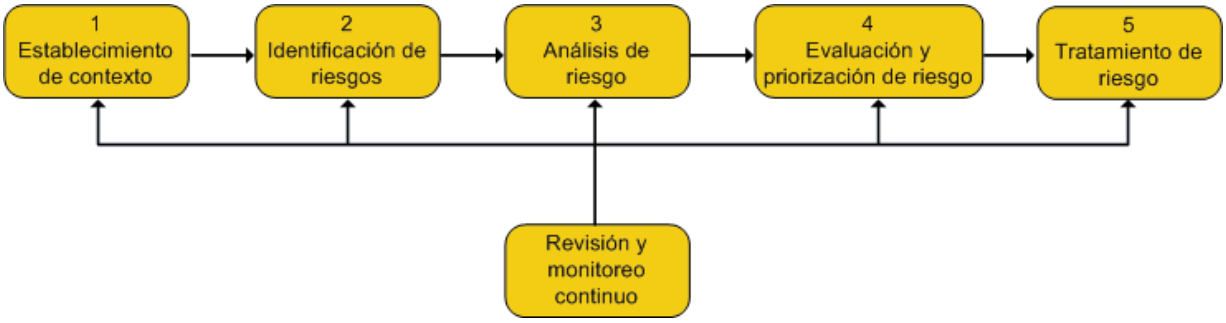

Figura 1: Proceso de gestión de riesgo.

Las tareas presentes en cada una de las etapas del proceso de gestión son las siguientes:

1. **Establecimiento del contexto:** Se entregan los objetivos, valores, políticas y estrategias que ayudan a definir los criterios que determinarán, finalmente, cuáles de los riesgos identificados son aceptables y cuáles no; así como establecer las bases de los controles necesarios y la administración de las opciones disponibles para enfrentarlos. El enfoque y profundidad de la revisión de riesgos también debe definirse en esta etapa.

2. **Identificación de riesgos:** Se identifican los riesgos y la relación de éstos con los controles existentes para posteriormente trabajar en base a los resultados obtenidos. Una buena identificación de los riesgos involucra examinar todas sus fuentes y las perspectivas de todos los entes participantes, ya sean internos o externos. Otro factor importante es la buena calidad de la información, y comprender cómo y dónde estos peligros han tenido o pueden tener su efecto.

3. **Análisis de riesgos:** Habiendo ya identificado los riegos, se analizan las posibilidades y consecuencias de cada factor de riesgo. Los peligros necesitan ser analizados para decidir cuáles son los factores de riesgos que potencialmente tendrían un mayor efecto y por lo tanto, necesitarían ser analizados o tratados. Existen tres categorías de métodos utilizados para determinar el nivel de riesgo, los cuales son:

    - **Cualitativos:** Un análisis cualitativo se puede utilizar cuando el nivel de riesgo no justifican el tiempo y recursos para realizar un análisis más detallado; también para obtener una observación general al inicio de los riesgos, que concluye en un análisis posterior más detallado.

    - **Semi-cuantitativos:** Un enfoque semi-cuantitativo puede utilizar clasificaciones de palabras como "alto", "medio" y "bajo", o bien descripciones más detalladas de la probabilidad y la consecuencia.

    - **Cuantitativos:** El nivel de riesgo se puede realizar utilizando el método cuantitativo en las situaciones donde la probabilidad de ocurrencia y las consecuencias puedan ser cuantificadas.

4. **Evaluación y priorización de riesgo:** En esta etapa se decide si los riesgos son aceptables o no. Esto se logra comparando el nivel de cada riesgo obtenido como salida de la etapa anterior, en relación con el nivel de riesgo aceptable evaluado en la etapa uno. La evaluación debe tomar en cuenta el grado de control sobre cada uno de los peligros, el impacto de costo, beneficios y las oportunidades presentadas por ellos. Finalmente, los riesgos se clasifican de acuerdo a las prioridades de gestión para su tratamiento.

5. **Tratamiento de riesgo:** El tratamiento de cada riesgo identificado tiene que ser adecuado a la significancia del mismo, y también de acuerdo a la importancia del programa, proceso o actividad asociada. Como pauta general se pueden mencionar que los riesgos:
    - De bajo nivel pueden ser aceptados por lo que no es necesaria una acción adicional.
    - De nivel medio corresponden a un nivel relevante, el cual tiene que ser tratado.
    - De alto nivel requieren de una cuidadosa administración o gestión y de la preparación de un plan formal para su administración.

6. **Revisión y monitoreo continuo:** Esta etapa es esencial e integral en el proceso de gestión de riesgo. Es necesario monitorear los peligros, la efectividad del plan, las estrategias y el sistema de administración establecido. Las situaciones peligrosas necesitan ser controladas periódicamente para garantizar que las circunstancias cambiantes no alteren las prioridades definidas.

La gestión de riesgo es un elemento fundamental para el SNA, ya que a partir de ella se priorizan los riesgos a tratar por medio de mecanismos de control como la fiscalización. Posteriormente, con los resultados que se obtienen en los procesos de fiscalización, se pueden identificar nuevos riesgos, los cuales son utilizados como parte de un flujo de retroalimentación que tiene por objetivo entregar la información necesaria para ajustar el modelo de gestión de riesgos.

2.1.3. Proceso de Fiscalización

Con el fin de comprender la integración entre el modelo de inteligencia y la gestión de riesgo, se presenta al proceso de fiscalización como un macro-proceso que los engloba. A continuación, se describen las distintas etapas de este macro-proceso [3]:

- **Obtención y tratamiento de la información**: Esta etapa tiene la responsabilidad de obtener la información necesaria del entorno activo, clasificarla, archivarla, velar por su seguridad y poner físicamente a disposición de todas las etapas del proceso, la información generada en cada una de éstas; además de administrar los recursos tecnológicos relacionados.
- **Análisis de inteligencia**: Esta etapa "procesa" la información producida y desarrolla los primeros análisis, principalmente de tipo económico, comercial o estadístico, histórico o proyectado, agregado o específico, con el objeto de generar información de apoyo para sustentar la etapa de identificación, análisis y evaluación de riesgos y las demás etapas que requieran información para su operación.
- **Identificación, análisis y evaluación de riesgo aduanero**: En esta etapa se realiza la selección de las áreas, sectores o agrupaciones de operaciones para finalmente identificar los agentes económicos que serán sometidos a una investigación. O sea, para cada nivel de estudio se aborda la problemática de la selectividad, en base a las consideraciones de riesgo aduanero (posibilidad de fraude, contrabando u otras infracciones aduaneras). Una vez identificado el riesgo se procederá a su tratamiento para su eliminación, disminución o traslado.
- **Diseño de la estrategia de tratamiento**: Una vez especificado un sujeto o eventualmente un área de riesgo, se desarrolla un proceso de diseño de la estrategia más adecuada para el tratamiento del riesgo identificado, en el cual se evalúa el grado de cumplimiento de la normativa aduanera que regula la situación bajo investigación.
- **Ejecución de tratamiento/recolección de evidencia**: En esta etapa se lleva a cabo la ejecución del tratamiento. La naturaleza de éste dependerá de las características del riesgo identificado y los objetivos de control. Así, si el riesgo se relaciona con un fraude comercial identificado en un grupo particular de empresas, el tratamiento tendrá forma de un programa de fiscalización de carácter coercitivo, probablemente aplicando técnicas de auditoría en dependencias de las empresas para recabar las evidencias que sustentan la existencia del ilícito. Si al contrario, el riesgo identificado se relaciona con debilidades de control, probablemente la estrategia adecuada sea el mejoramiento de las normas reglamentarias lo que se ejecutará por medio de una propuesta de mejoramiento.
- **Dirección**: Esta etapa es responsable de la gestión general del proceso, dictando las políticas y definiendo las directrices estratégicas para el funcionamiento y desarrollo del sistema.
- **Planificación y control de gestión**: Se encarga del diseño de los planes y programas de la actividad fiscalizadora, es decir, de la estrategia a seguir para el logro de los objetivos establecidos. Considera además

el control y la evaluación del desempeño, y en general de la administración de un sistema de planificación y control de gestión.

- **Coordinación**: Esta instancia es responsable de hacer que los enfoques, ritmos, esfuerzos o intereses individuales, sean funcionales a las metas y objetivos organizacionales.
- **Mejoramiento continuo**: Todo el proceso se encuentra inserto dentro de un marco de mejoramiento continuo, el cual permite optimizar los procesos por medio de la evolución de su desempeño.

2.1.4. Temporización de la Fiscalización

Para estudiar el proceso de fiscalización, éste se enmarcará sólo en el proceso de importación, el cual tiene asociado el documento llamado Declaración de Ingreso, conocido como DIN. Se ha optado por trabajar sólo a nivel de importaciones con el fin de disminuir la complejidad del estudio a realizar, sin embargo, los resultados que se obtengan pueden ser aplicados tanto a nivel de importaciones como exportaciones, donde el documento utilizado recibe el nombre de Declaración Única de Salida (DUS).

Desde el punto de vista de la temporización para la fiscalización de las importaciones se pueden diferenciar tres tipos, en función del instante en el cual se lleva a efecto la fiscalización, la que puede ser física o documental, tal como lo muestra la Figura 2. Estos tres tipos corresponden a:

1. **Fiscalización a priori:** Corresponden a las actividades asociadas a la validación a nivel documental, del contenido que figura en el documento de ingreso. En esta instancia también se encuentran los mecanismos de control en base a filtros, los cuales se ejecutan en forma automatizada una vez que se recibe la información asociada a la importación, por medio de documentos DIN electrónicos.

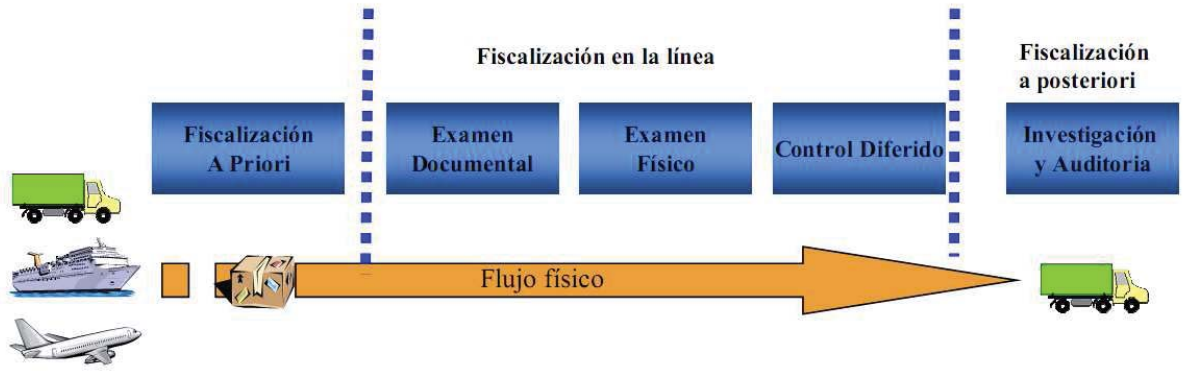

Figura 2: Etapas de fiscalización

2.- Fiscalización en la línea: El proceso de fiscalización "en la línea" (no confundir con fiscalización en línea), corresponde al conjunto de operaciones que se pueden realizar en un punto de control donde se tiene a mano la documentación física y las mercancías. Este proceso se puede describir de la siguiente forma:

- **Verificación de la información:** La información recibida es verificada para asegurar el cumplimiento de aspectos como: normas aduaneras, arancel aduanero y cupos.
- **Selección para revisión:** Después de realizar las verificaciones anteriores, se procede a la selección de las operaciones que serán fiscalizadas. La selección se lleva a cabo de tres maneras distintas: mediante el sistema de selección de aforo (utilización de filtros), por decisión del fiscalizador en zona primaria y mediante los resultados obtenidos por sistemas de selección de carácter predictivo.
- **Aforo:** Es el proceso de revisión de las mercancías. Éste puede ser físico o documental.
- **Resultados del aforo:** Como resultado del aforo se pueden presentar las siguientes alternativas: todo en orden (el proceso sigue su curso normal), el fiscalizador abre una carpeta de investigación para que sea fiscalizada a posterior, el fiscalizador decide realizar una toma de muestras y por último, si se detecta una operación irregular, puede terminar en la generación de una denuncia.

2. **Fiscalización a posterior:** La fiscalización a posterior es completamente "documental" puesto que en la mayoría de los casos cuando se revisa la mercancía, ésta ya no se encuentra en poder del agente económico respectivo; por lo mismo, la fiscalización está orientada al agente económico y no a la mercancía. Un aspecto relevante en este tipo de fiscalización corresponde al período de tiempo en el cual se puede aplicar, ya que a diferencia de las anteriores, que son de carácter inmediato o en plazos de tiempos muy acotados, la revisión a posteriori puede ser aplicada a transacciones cuya antigüedad puede llegar hasta los 3 años.

Por razones asociadas a la simplificación de los temas tratados, se ha optado por sólo trabajar con la Fiscalización a Posteriori*,* dejando fuera del alcance de esta tesis a las fiscalizaciones restantes, ya que por tratarse de una etapa experimental, resulta apropiado trabajar en un entorno flexible en cuanto a la disponibilidad de tiempo, escenario que claramente sólo se encuentra disponible en este último tipo de fiscalización.

2.1.5. Herramientas de Apoyo Para las Tareas de Inteligencia

La naturaleza y características de los procesos de control, han tenido que evolucionar en la medida en que también lo ha hecho, cualitativa y cuantitativamente, el comercio internacional. Dentro de los factores que han incidido notablemente en este nuevo escenario, se puede destacar la modernización del sector privado, donde se han informatizado los procedimientos y optimizado el transporte de las mercancías, sumado al empuje que ha recibido el comercio exterior producto del creciente aumento de tratados de libre comercio y la notable masificación que está adquiriendo el comercio electrónico (e-commerce).

Sólo a nivel cuantitativo, la cantidad de importaciones anuales está superando el millón de transacciones (ver Figura 3), lo que contrasta notoriamente con las 300.000 operaciones promedio que se producían durante la década de los 80'[1]. Pero la complejidad de este nuevo escenario no radica sólo en el aumento cuantitativo de las importaciones y

[1] Según estadísticas del Servicio Nacional de Aduanas

exportaciones; los cambios propios de la evolución de la sociedad han incorporado un conjunto de nuevos desafíos que aumentan aún más los niveles de complejidad existentes. El surgimiento de nuevos retos asociados a ilícitos como el lavado de dinero, el narcotráfico, contrabando de sustancias consideradas peligrosas para el medio ambiente y la seguridad personal o nacional, traen consigo el surgimiento de nuevas responsabilidades.

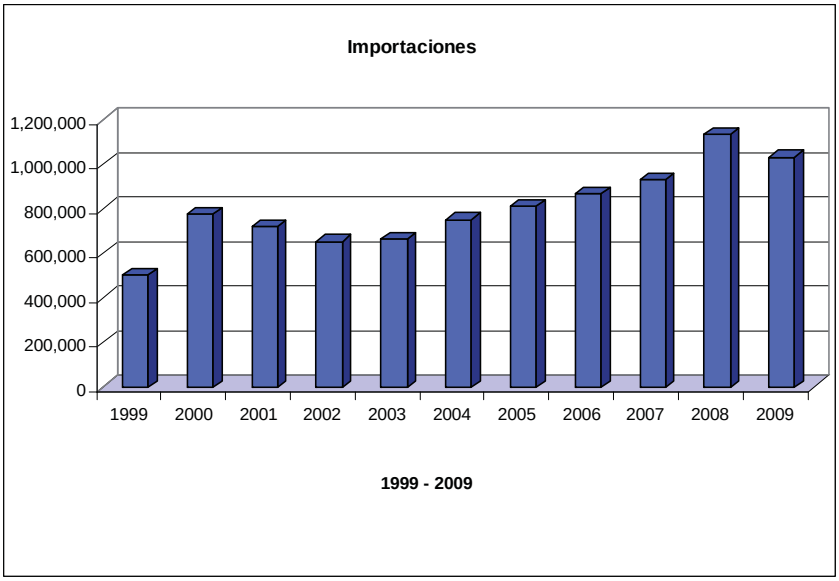

Figura 3: Importaciones anuales periodo 1999 – 2009.

Para enfrentar el nuevo escenario, el Estado destinó los recursos necesarios para impulsar un proceso de modernización del servicio focalizado en las siguientes objetivos críticos [2]:

- Modernizar el sistema de fiscalización propiamente tal, pasando de un sistema aleatorio y principalmente físico, a uno basado en la evaluación de riesgo y centrado en el análisis a posteriori de las mercancías.
- Actualizar la plataforma tecnológica del SNA para apoyar de mejor forma los procesos, con el fin de disminuir los tiempos de respuesta.
- Fomentar el desarrollo de las relaciones con las distintas áreas de fiscalización el aparato público (Servicio de Impuestos Internos, Servicio Agrícola y Ganadero, etc.).
- Mejorar las instalaciones para ampliar el horario de atención.

Lo anterior tenía por objetivo mejorar la fiscalización para poder enfrentar de manera apropiada el notable aumento de las transacciones de comercio exterior, el cual llevaba implícito un aumento de la complejidad de las mismas. Es evidente que no es posible fiscalizar todo lo que entra y sale del país, más aún, si la cantidad de recursos destinados para esta tarea no da abasto para manejar el crecimiento explosivo de los últimos años. Por esta razón, el cambio de enfoque aplicado al proceso y la incorporación de tecnología para potenciar las transacciones electrónicas parecían ser el camino correcto.

El cambio anterior fue la base para mejorar los procesos de recepción y validación por medio de controles automatizados, lo que disminuyó notablemente los tiempos de respuesta involucrados en las tramitaciones. Esto último, también fue un incentivo para que los agentes de Aduana adoptaran rápidamente la tramitación electrónica, lo que agilizó el proceso de migración a este nuevo escenario, en el cual se pretendía mejorar la fiscalización, sin que esto último entorpeciera el normal desarrollo del comercio exterior.

Después de transcurrida más de una década desde que comenzó la modernización del SNA, se puede afirmar que se cuenta con el escenario propicio para incorporar herramientas que, a partir de la información histórica de las transacciones que registra Aduana, puedan realizar análisis sobre los datos para obtener el conocimiento subyacente. En un modelo donde el análisis de riesgo y la inteligencia juegan un papel fundamental para guiar y optimizar los resultados de la fiscalización, la utilización de herramientas que apoyen dichas tareas de inteligencia resultan ser de vital importancia para identificar los comportamientos y tendencias que se encuentran ocultas en las propias transacciones.

2.2. ÁMBITO DE TESIS

Los temas abordados en la sección anterior tenían por objetivo realizar una breve reseña sobre conceptos básicos asociados al negocio, así como también, plantear los desafíos a los cuales se ve enfrentado el Servicio y la evolución que éste ha presentado desde el inicio de su modernización.

En base a lo antes descrito, se cuenta con los elementos necesarios para proceder a delimitar el dominio de acción en el cual se enmarcará la solución a definir, la que considera:

- **Proceso de negocio**: El estudio se realiza sobre el proceso de fiscalización asociado a las importaciones, encargándose de revisar las transacciones de los documentos de importación DIN para realizar una clasificación sobre estas, con el fin de detectar documentos declarados en forma irregular, que podrían estar asociados a un comportamiento fraudulento.

- **Temporalidad**: A pesar que el SNA históricamente ha invertido recursos para disponer de una plataforma de clasificación en base a modelos [1,4,5], a la fechas estos esfuerzos se mantienen como soluciones experimentales. Por esta razón no se recomienda este método para apoyar a la fiscalización a priori o en línea, sin antes alcanzar la madurez adecuada en un entorno más flexible. Por lo anterior, la estrategia adoptada es, primero apoyar al proceso de fiscalización a posteriori y aprender de la experiencia alcanzada en esta modalidad, para una vez alcanzado un nivel de estabilidad y rendimiento adecuado, proceder a incluir los procesos online, donde el tiempo de respuesta es fundamental.

- **Tipo de fraude**: El tipo de fraude que se trata de controlar es de derechos dejados de percibir por subvaloración de mercancías.

- **Fiscalización adecuada**: Los procesos automatizados para detectar fraude tienen que detectar casos

riesgosos, sin que la ejecución de dichos procesos llegue a entorpecer las operaciones de comercio exterior. En otras palabras, los procesos deben de presentar un rendimiento que permita ejecutar adecuadamente las tareas de inspección y clasificación, sin producir retrasos excesivos en el flujo de las operaciones de importación.

- **Área del negocio**: El departamento para el cual está destinada la herramienta a desarrollar corresponde al área de inteligencia aduanera.
- **Riesgo**: La solución propuesta tiene que estar relacionado con los mecanismos de gestión de riesgo que posee el área de fiscalización.

3. ESTADO DEL ARTE

En este capítulo se realizará un recorrido sobre el estado del arte de la minería de texto y de datos. Lo primero que se define son aquellos conceptos básicos, algunos de los cuales son elementos comunes para los dos tipos de análisis de datos; se han seleccionado aquellos que se consideran fundamentales para poder comprender las técnicas que se usaran posteriormente en la solución planteada. Un tratamiento especial reciben las técnicas de minería de texto, ya que los principales problemas a resolver en las etapas de pre-procesamiento y transformación de los datos recaen en estas técnicas.

En lo que respecta a minería de datos, se profundizará en los aspectos asociados al estado del arte para decantar en las técnicas asociadas a la detección de anomalías basadas en la identificación de *outliers* por medio de agrupamientos.

Además, como este trabajo se encuentra inmerso en el ámbito de la detección de fraudes, se realiza una breve reseña de este tema, que tiene por objetivo mostrar cómo las técnicas de minería se pueden utilizar para la detección de fraude.

3.1. CONCEPTOS BASICOS

Tanto la minería de texto como de datos se basan en recursos interdisciplinarios, los cuales son utilizados en conjunto para apoyar a las tareas que se encargan de la extracción del conocimiento. Casos especiales, como las áreas de inteligencia artificial y la teoría de la información, resultan ser de principal interés ya que son ampliamente utilizados en todo proceso de minería que ocupa técnicas automatizadas, indistintamente si se trata de texto o datos. Por lo tanto, resulta necesario dominar algunos conceptos básicos relativos a estas áreas, con el fin de proveer la base conceptual necesaria para comprender el funcionamiento de las herramientas que serán utilizadas en la confección del *framework*.

Dado lo anterior, se procede primero a realizar una breve reseña sobre conceptos relevantes asociados a lingüística, teoría de información, y procesamiento de lenguaje natural (PLN).

3.1.1. Lingüística

Antes de abordar las técnicas de minería de texto, es necesario comprender los principios teóricos asociados al lenguaje. Por esta razón, resulta fundamental realizar un pequeño recorrido que contemple algunos fundamentos de lingüística.

En lo que respecta a teoría lingüística, el análisis del lenguaje se realiza en base a una estructura estratificada, la cual comienza con unidades de lenguaje básicas, las que a su vez aumentan su complejidad lingüística a medida que se asciende a los niveles superiores de una estructura estratificada, donde cada capa corresponde a un nivel lingüístico distinto [8]. De este modo, los diferentes sonidos utilizados en el lenguaje se pueden encasillar en el nivel *Fonológico.* El estudio de los sistemas escritos los podemos asociar a un nivel que contempla la Ortografía. La

Morfología, se encarga de la estructura de las palabras donde se considera la forma de estas y las inflexiones que puedan presentar.

La Sintaxis describe el ordenamiento de las palabras y cómo éstas pueden ser combinadas para formar estructuras más complejas como frases y sentencias. La Semántica se encarga de analizar el significado de las palabras individuales y de éstas en unidades más complejas como frases y sentencias, recibiendo el nombre de Semántica Composicional. El nivel Pragmático se encarga de analizar cómo las palabras y frases se relacionen en un contexto. Y por último, el nivel de discurso corresponde a la forma en que las personas y las cosas son presentadas en forma de temas y subsecuentemente nombradas en declaraciones.

Cada una de las descripciones lingüísticas mencionadas, pueden ser consideradas como capas que representan distintos niveles de complejidad en la estructura del lenguaje, tal como lo muestra la Tabla 1.

Niveles	Descripción
Discurso	Cohesión en texto y dialogo
Pragmático	Funciones de expresiones
Semántico	Significado de palabras y sentencias
Sintáctico	Orden de palabras y estructura de sentencias
Morfológico	Formación de palabras e inflexiones
Ortográfico	Lenguaje escrito
Fonológico	Sonidos (lenguaje hablado)

Tabla 1: Niveles de estudio para el lenguaje.

Es importante aclarar que para el tratamiento del lenguaje no necesariamente se tienen que cubrir todos los niveles o seguir estrictamente el orden establecido. Por ejemplo, para el desarrollo de la tesis sólo se consideran los niveles que tienen relación con el lenguaje escrito, seleccionado sólo aquellos contemplados entre la ortografía y la semántica, inclusive. La justificación de los niveles a incluir en el análisis tiene relación directa con la complejidad del texto fuente y el nivel de análisis necesario para obtener los resultados esperados.

Por otro lado, en base a la literatura existente, se tiene que estos niveles pueden variar ya que su inclusión dependerá del grado de profundidad en aspectos lingüísticos que el autor haya decidido incorporar en su obra, así como también, limitar la incorporación de sólo aquellas capas que tienen relación con el alcance de los contenidos tratados. Además es importante tener presente que también pueden existir algunas sutiles diferencias en los nombres asignados a éstos, pero el sentido de fondo es el mismo, tal como se puede apreciar en [13,14,15].

3.1.2. Teoría de la Información

Es importante aclarar que en lo que respecta a la teoría de la información existían trabajos previos, como los desarrollados por Norbert Wiener [11,12], donde se plantea el concepto de entropía asociado a la información. Sin embargo, es Shannon, ingeniero electrónico y matemático, quien se preocupa de realizar una definición formal de

información [9]; algunos de los aspectos sobresalientes de su teoría serán tratados a continuación con el fin de recordar conceptos fundamentales que son la base de la teoría de la comunicación moderna.

Entropía

La entropía, en el plano de la información, puede ser definida como carencia de contenido de información en un mensaje o como un estado de desorden de la misma. Se debe aclarar, que por lo general se acostumbra definir la información como un conjunto de datos útiles, pero en teoría de la información, ésta también es incertidumbre. Es decir, la información ayuda a disminuir la incertidumbre, por lo tanto es un flujo neguentrópico, pero si ésta aumenta su estado de desorden, producirá el efecto contrario, aumentando la entropía.

Para obtener la fórmula de la entropía, primero se define el bit como unidad de información. De este modo, si se considera que con un bit se puede determinar un valor booleano, se puede tener *Verdadero* =1 y *Falso* =0.

Si se tiene que los posibles valores para el atributo v_i ocurre con una probabilidad igual a $P(v_i)$, entonces el nivel de entropía estará dado por:

$$E(P(v_i),...,P(v_n)) = \sum_{i=1}^{n} - P(v_i)\, Log_2 P(v_i) \qquad (1)$$

donde *E* representa al nivel de entropía, $P(v_i)$ es la probabilidad de que un evento v_i ocurra y *N* es el número total de eventos.

Si se utiliza la función logaritmo en base 2, se tendrá que la entropía se está midiendo en bits. Para aclarar mejor el concepto se tiene el siguiente ejemplo: considerar el ejemplo lanzamiento de una moneda al aire; la probabilidad de que salga cara o sello será de un 50% para cada caso. Utilizando la fórmula (1) se tiene:

$$E(cara, sello) = -P(cara)Log_2 P(cara) - P(sello)Log_2 P(sello) = E\left(\frac{1}{2}, \frac{1}{2}\right)$$

$$E\left(\frac{1}{2}, \frac{1}{2}\right) = -\frac{1}{2} Log_2 \frac{1}{2} - \frac{1}{2} Log_2 \frac{1}{2} = 1$$

De lo anterior, se obtiene que en una distribución uniforme todos los valores son igualmente probables y por lo mismo la entropía es máxima (igual a 1), lo que lleva a un máximo nivel de incertidumbre.

Por otro lado, si se tiene la certeza absoluta de que siempre saldrá cara, la fórmula cambia a:

$$E(1,0) = -1Log_2 1 - 0Log_2 0 = 0$$

Para este caso en donde la $P(v_1)$=1 y $P(v_2)$=0, el nivel de entropía es el mínimo. Esto implica, mínima incertidumbre, lo que se traduce en una mayor información.

La Figura 4 muestra la gráfica de la curva de entropía obtenida para el ejemplo del lanzamiento de la moneda.

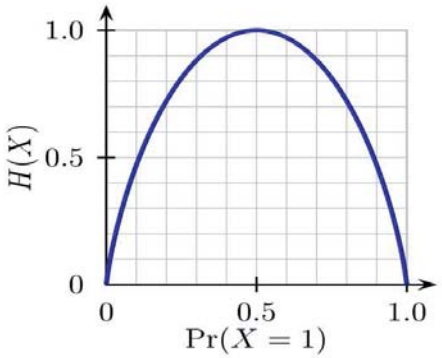

Figura 4: Gráfica de la función entropía para valores booleanos.

Ganancia

Por otro lado, la *ganancia* se define como: restar a la entropía global la media ponderada de las entropías que pueda tomar un atributo.

$$\text{Ganancia}(S, A) = E(S) - \sum_{v \in A} \frac{|S_v|}{S} E(S_v)$$

Varios de los temas asociados a la utilización de probabilidades para procesamiento de lenguaje, donde también se explica el concepto de entropía en teoría de la información, son mencionados en [7].

Información Mutua (MI)

Corresponde a otro concepto importante a tener presente, ampliamente utilizado en probabilidades y teoría de la información. Una reseña detallada así como su aplicación a la identificación de concurrencias en texto puede ser consultada en [7].

La información mutua entre dos palabras o tokens w_1 y w_2 , se define de la siguiente forma:

$$MI(w_1, w_2) = Log_2 \frac{P(w_2 \mid w_1)}{P(w_2)}$$

La medida de información mutua entre dos palabras es una herramienta útil en minería de texto, para determinar si existe una fuerte relación entre palabras que frecuentemente aparecen juntas. Cuando se realiza un análisis de

coocurrencias sobre palabras, es posible identificar las relaciones que sobresalen, las que reciben el nombre de collocation.

Para identificar relaciones entre un conjunto de palabras que aparecen frecuentemente juntas, basta con realizar un análisis de frecuencia. Sin embargo, cuando la complejidad suele ser mayor, es recomendable recurrir a medidas más complejas como el cálculo de información mutua.

3.1.3. Procesamiento de Lenguaje Natural

Lenguajes naturales y formales

Las lenguas son sistemas que permiten manifestar pensamientos por medio de expresiones simbólicas, tanto orales como escritas; así es posible definir al lenguaje como la capacidad humana de comunicarse mediante lenguas. Este concepto es asociado a los lenguajes naturales, aquellos que surgen sin una teoría que los genere; también se aplica a lenguajes creados por el hombre, los cuales reciben el nombre de lenguajes formales (lógica, matemática, lenguajes de programación).

Formalizando lo anterior, se tiene que en un lenguaje se pueden apreciar aquellos símbolos más básicos llamados letras, que constituyen un alfabeto denotado por $\sum$, que se caracteriza por ser un conjunto finito. Con la concatenación de los símbolos se pueden formar palabras las cuales determinan un conjunto llamado $\sum^*$. El conjunto de palabras que tenga un significado corresponderá al diccionario. En resumen, un lenguaje corresponde a un conjunto de oraciones las cuales pueden ser infinitas, por la característica recursiva del lenguaje, las que a la vez son formadas por las palabras del diccionario.

En los lenguajes naturales la formación de los términos y las oraciones precede a la formación del lenguaje por medio de una teoría o gramática; es por esto que reciben el nombre de natural, asociado a su naturaleza, donde no se aplica una generación artificial. Además, en este tipo de lenguajes las palabras tienen un significado que puede variar según el contexto semántico. Esto da origen a la ambigüedad, un problema fundamental sobre el cual se han realizado múltiples estudios y propuestas para manejarlo, como parte de las operaciones importantes dentro del procesamiento de lenguaje.

En resumen, los lenguajes naturales contienen las siguientes propiedades:

- Son formados en base a enriquecimiento progresivo, el cual no se basa en una teoría.
- Su carácter expresivo se relaciona directamente a su riqueza semántica.
- Por su naturaleza es muy difícil lograr una formalización completa.

Por otro lado, a diferencia de los lenguajes naturales, en los formales existe una teoría que precede a su definición. En éstos, las palabras y oraciones tienen siempre el mismo significado, el cual no depende de un contexto semántico; además, el significado de los símbolos utilizados es determinado exclusivamente por la sintaxis de los mismos. Por

lo anterior, los lenguajes formales carecen de cualquier componente semántico más allá del alcance definido por sus operadores y relaciones.

Resumiendo, se tiene que los lenguajes formales se caracterizan por los siguientes aspectos:

- Son desarrollados a partir de una teoría.
- Tienen una cantidad limitada de componentes semánticos mínimos.
- Su sintaxis permite obtener oraciones no ambiguas.
- Poseen una formalización completa.

Como dato adicional, es posible profundizar sobre los aspectos asociados al lenguaje y la gramática consultando [15].

Modelos de lenguaje

Un modelo de lenguaje corresponde a una representación matemática, específicamente un modelo estadístico, cuyo objetivo es predecir la secuencia de las palabras. Dentro de los modelos de lenguaje más conocidos se encuentra el de N-gramas .Para generalizar, se puede decir que un N-grama es una secuencia de ítemes los cuales pueden ser tan variados como: sílabas, letras y palabras.

Para clarificar lo anterior, se presenta un ejemplo de Bigrama y Trigrama a nivel de palabras sobre un texto. Las salidas que a continuación se presentan, fueron obtenidas utilizando la implementación de N-gramas provista por la API NLTK, cuya ficha técnica se puede encontrar en la sección 4.2.

Texto : “Procesamiento de lenguaje natural”

Bigrama: [('procesamiento', 'de'), ('de', 'lenguaje'), ('lenguaje', 'natural')]

Trigrama: [('procesamiento', 'de', 'lenguaje'), ('de', 'lenguaje', 'natural')]

También es posible utilizar un autómata finito para representar los lenguajes generados por unigramas, bigramas y trigramas, tal como se puede apreciar en la Figura 5.

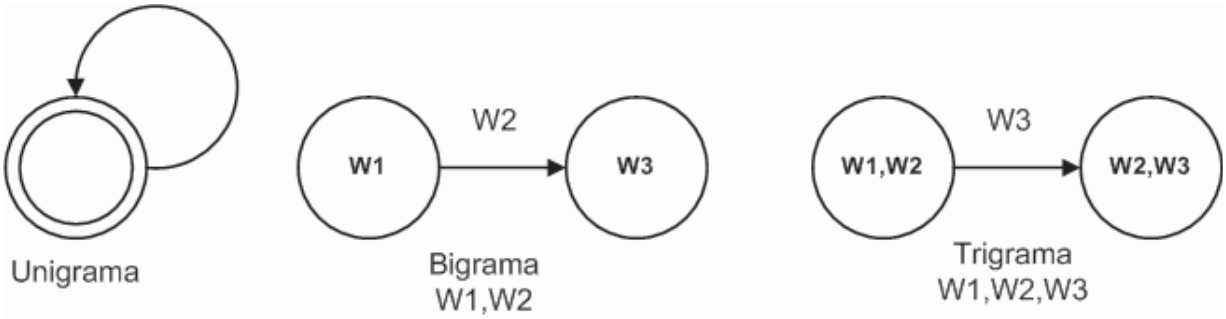

Figura 5: Unigramas, bigramas y trigramas como grafo.

Un modelo de lenguaje basado en N-gramas se preocupa de utilizar la historia de los N-1 elementos inmediatos que preceden para calcular la probabilidad *P* de la N-ésima palabra. En general, los tamaños de los N-gramas más comunes son los que se han mencionado anteriormente, esto debido a que generar un modelo en base a N-gramas aumenta su costo a medida que crece el tamaño de *N*. Por consiguiente, un modelo de lenguaje basado en N-gramas típicamente lista sólo las más frecuentes ocurrencias de palabras para una ventana de un tamaño N razonable.

Las probabilidades para los N-gramas se obtienen con la siguiente formula:

$$P(w_1,\ldots,w_{n-1}) = \prod_{i=1}^{n} P(w_i,\ldots,w_{i-1})$$

En un modelo de lenguaje en base a N-gramas, aquellas palabras o conjunto de estas que aparecen frecuentemente en un texto recibirán una alta puntuación en base a su frecuencia, mientras que las que menos aparecen recibirán una puntuación menor. Adicionalmente, existen aquellos casos especiales que corresponden a las palabras que no figuran en el modelo, las cuales tendrán una probabilidad igual a cero, ya que son desconocidas en el modelo actual. Como contar con valores de probabilidad igual a cero puede ser muy extremo en el modelo, surge la necesidad de recurrir a alguna transformación que "suavice" estos valores. Dicho suavizado recibe el nombre de *smoothing*, donde uno de los mecanismos más simples corresponde a asumir que estas palabras nuevas se han visto a lo menos una vez.

Los modelos basados en N-gramas, ya sea de palabras o letras, son ampliamente utilizados para obtener estadísticas de utilidad para ser ocupadas en un proceso de PLN. También se usan frecuentemente en reconocimiento de lenguaje hablado donde los fonemas y las secuencias de estos pueden ser modelados en base a N-gramas. Además se utilizan para el reconocimiento de lenguajes, donde la representación de secuencias de letras permite determinar el lenguaje utilizado.

Existe una infinidad de áreas donde la necesidad de obtener un comportamiento estadístico de los elementos presentes en un documento, resulta de apoyo para realizar tareas de procesamiento de texto, donde a las ya comentadas anteriormente podemos agregar las siguientes:

- Determinar cuáles son los candidatos más recomendables para una palabra que presenta error, en aplicaciones de corrección ortográficas.
- Mejorar el rendimiento de algoritmos de compresión.
- Mejorar la exactitud en la recuperación de documentos en un sistema de recuperación de información.
- Mejorar el rendimiento de sistemas que trabajan con análisis de secuencias genéticas.

Los temas asociados a modelos de lenguaje basados en N-gramas se pueden consultar en [7, 13, 14,56].

Herramientas utilizadas en PLN

En lo que respecta al tratamiento del lenguaje, los niveles antes mencionados en la Tabla 1 son fundamentales para delimitar las tareas que se llevan a cabo. Las técnicas que se aplican se caracterizan por ser especializadas en un nivel

determinado, pero esto último no implica la ausencia de comunicación entre capas ya que muy por el contrario, es común ver que los resultados obtenidos en el procesamiento de un nivel específico, son utilizados como entrada por las tareas de un nivel superior. Esta forma cooperativa de trabajo entre los procesos que se enmarcan dentro de los distintos niveles de análisis para el lenguaje, corresponde al enfoque tradicional utilizado en los sistemas de PLN, donde la comunicación se puede dar a nivel de procesos de una misma capa, entre procesos de capas distintas (generalmente ascendente entre capas adyacentes) y entre sistemas de procesamiento de texto, lo que recibe el nombre de *pipeline*.

Dentro de las tareas más características asociadas a cada nivel se encuentran las detalladas en la Tabla 2, donde se plantea una abstracción de aquéllas que se encuentran disponibles en base al estado del arte en esta área. Para cada tarea existen una amplia variedad de implementaciones que utilizan distintas técnicas con el fin de mejorar los resultados que se puedan obtener. Sin embargo, es común para cada procesamiento, independiente del nivel en el cual trabaje o de las técnicas utilizadas, el apoyarse de un sistema de anotaciones, las cuales se agregan al texto procesado por medio de un etiquetado (Labels o Tags).

Procesamiento	**Descripción**
Resolución de conferencia	Asocia referencias con alguna entidad en el texto.
Reconocimiento de entidades	Identifica y etiqueta los nombres de las entidades
Análisis semántico	Analiza la relación predicado-argumentos
Parsing sintácticos	Analiza las frases que forman una sentencia
Tokenización	Lenguaje escrito
Etiquetador POS	Etiqueta tokens con categorías de palabras
Frontera de sentencias	Segmenta el texto en sentencias

Tabla 2: Tareas de procesamiento de texto.

En la Tabla 2 el procesamiento de la fonología se encuentra descartado de la tabla ya que no se encuentra dentro de los alcances de este trabajo. En la tabla se comienza con la identificación de las fronteras entre las sentencias, acompañada por la tokenización del texto; estas tareas se pueden enmarcar dentro del nivel ortográfico y se encargan de obtener una segmentación a nivel de tokens, los cuales generalmente corresponden a palabras. Después de que el texto se encuentra segmentado en tokens, éstos pueden ser etiquetados con la categoría que le corresponde a cada palabra por un algoritmo del tipo Part of Speech (POS), el cual se encarga de identificar las partes de la oración como nombres, verbos, adjetivos, etc. Las tareas de tipo POS están asociadas al nivel Lingüístico ya que en base a la estructura de las palabras se puede identificar la categoría de estas; por ejemplo en la lengua hispana los verbos pueden terminar en "ar", "er" e "ir"; de este modo, dicha característica morfología puede ser utilizada por un parser del tipo POS para identificar los verbos en un texto en español.

Después que el texto está dividido en sentencias, éstas pueden ser analizadas como elementos que constituyen frases. Esta tarea de análisis recae en los parsers *sintácticos*, los cuales establecen una clara correspondencia entre el *nivel* sintáctico y las anotaciones que generan. En este nivel la complejidad del tratamiento del texto comienza a aumentar,

ya que es necesario recurrir a un conjunto de teorías lingüísticas, las cuales pueden ser complementadas por una amplia variedad de enfoques disponibles para el procesamiento sintáctico.

Para los niveles lingüísticos superiores (semántico, pragmático y discurso) se dispone de una enorme cantidad de técnicas propuestas, pero no existe un consenso sobre cuáles sobresalen sobre las demás, cuando se extiende el ámbito del problema de un entorno teórico al práctico [16].

3.2. TAXONOMÍAS DE TÉCNICAS

Tanto la minería de texto como de datos comparten un conjunto de técnicas en común las cuales pueden ser aplicadas directamente en las etapas de procesamiento de los datos o ser parte del diseño de herramientas para el procesamiento de éstos. La Figura 6 presenta un esquema de las técnicas más sobresalientes, dentro de las cuales están incluidas la mayoría de las presentadas en [83], complementando lo anterior mediante un esquema expuesto en [51]. En la figura se destacan aquellas técnicas que serán utilizadas en el desarrollo del *framework*.

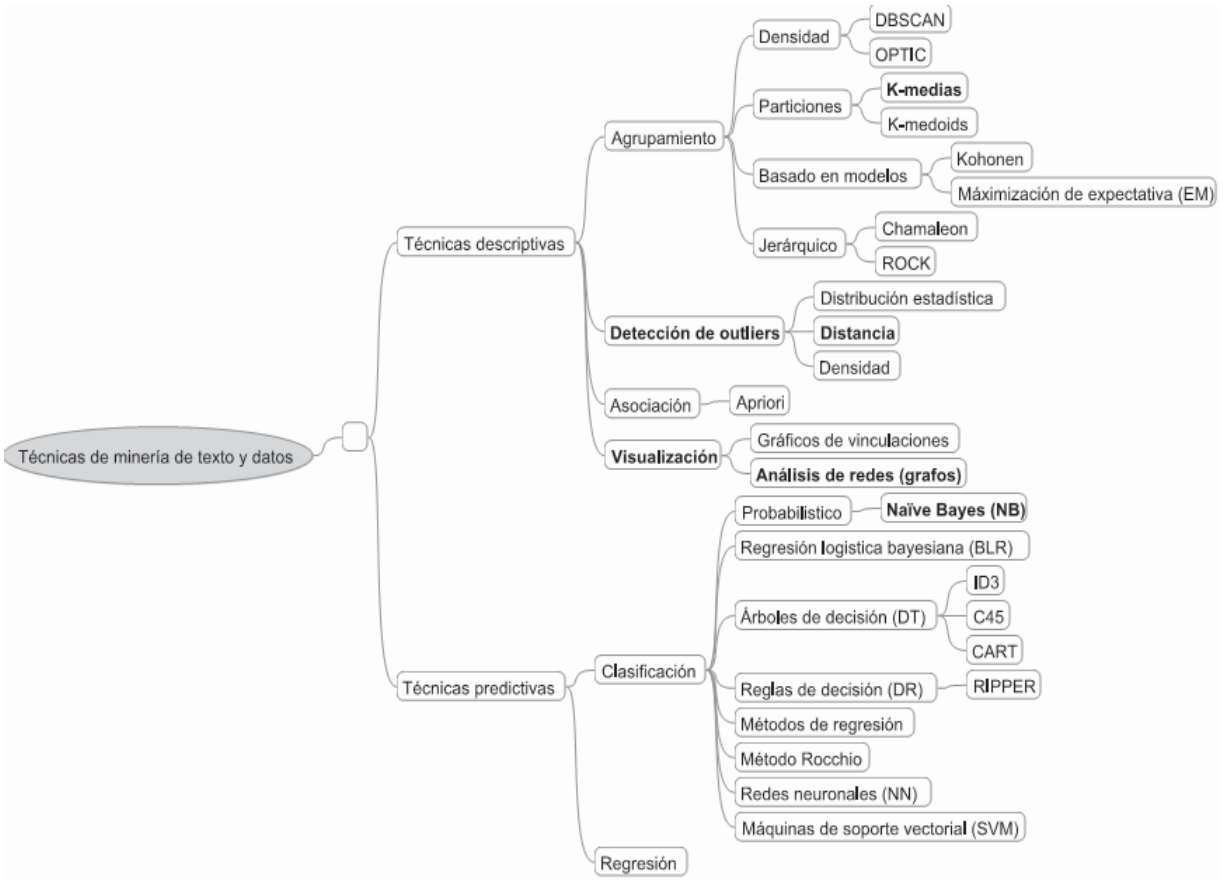

Figura 6: Taxonomía de técnicas aplicadas en minería de texto y datos

A continuación se procede a describir sólo algunos algoritmos que forman parte del esquema expuesto en la Figura 6. Principalmente aquellos que se encuentran más relacionados con los posibles enfoques a utilizar para enfrentar los desafíos planteados en este trabajo.

1. Agrupamiento *(*Clustering)

Es una técnica que permite distribuir los objetos en grupos, donde los componentes de cada uno son similares de acuerdo a alguna métrica. Los algoritmos de agrupamiento se pueden dividir en las siguientes categorías:

- **Particionamiento.** Se encargan de realizar *k* particiones sobre un conjunto de datos de tamaño *n*, con un *k* conocido y con $k<n$. En los algoritmos basados en particionamiento se tiene que cumplir que: (1) cada grupo tiene que tener a lo menos un objeto (2) cada objeto tiene que pertenecer exactamente a un grupo. Dentro de los algoritmos más conocidos se destacan *K-means* y *K-medoids.* [83,99]. El algoritmo K-mean se basa en clusters de prototipos, en términos de un centroide que corresponde a la media de un conjunto de puntos; típicamente se aplica a objetos en espacios continuos n-dimensionales. En el algoritmo K-Medias, al iniciar el proceso es necesario definir la cantidad de agrupaciones que se desean formar[83,99].
- **Jerárquicos.** Estos métodos de agrupamientos se preocupan de generar agrupaciones jerárquicas que son representadas como un árbol de clusters. Estos algoritmos pueden ser clasificados como *aglomerativos* o *divisivos*, dependiendo si realizan una clasificación jerárquica en base a un enfoque *botton-up* o *top-down.* Algunos algoritmos dentro de esta categoría son: DIANA, AGNES y BIRCH.
- **Basados en densidad.** Estos se encargan de realizar agrupamientos de regiones donde la población de los datos es densa, separados estos por regiones de baja densidad. Dentro de los algoritmos más conocidos se pueden destacar a DBSCAN y OPTIC [99].

2. Naïve Bayes

Se centra en la utilización de la teoría de probabilidad de Bayes, para estimar la probabilidad de que un objeto pertenezca a una clase. En el caso de minería de texto se utiliza para clasificar palabras dentro de una categoría la cual puede estar asociada a un tipo de documento. Un aspecto interesante de este clasificador, es que asume la independencia de las palabras, es decir, que la probabilidad condicional de una palabra dada una categoría, se calcula asumiendo que dicha palabra es independiente de las probabilidades condicionales de las palabras en el documento. A pesar que la aseveración se hace es falsa en la mayoría de los casos, el clasificador Naïve Bayes se desempeña mejor que otros [39, 51, 77]. A continuación se muestra la representación formal del teorema de Bayes.

$$P(A_i \mid B) = \frac{P(B \mid A_i)P(A_i)}{P(B)} = \frac{P(B \mid A_i)P(A_i)}{\sum P(B \mid A_j)P(A_j)}$$

donde $P(A_i)$ representa la probabilidad de cada una de las palabras en el documento y $P(B)$ es la probabilidad de la categoría.

3. Detección de anomalías

Se basa en encontrar objetos que presentan un comportamiento distinto a la tendencia de la mayoría, los que reciben el nombre de *outliers*. La detección de anomalías también se conoce como detección de desviaciones, ya que los objetos con anomalías tienen una desviación significativa respecto de los valores esperados.

En operaciones tradicionales de clustering, donde sólo se pretende agrupar a objetos con características similares, los *outliers* son considerados como "*ruido*" o errores, por lo que terminan siendo eliminados del conjunto de datos de trabajo. Sin embargo, en las operaciones de detección de fraude, es de vital importancia detectarlos y analizarlos; por esta razón en estos casos el análisis se centra en la identificación y estudio de *outliers* para encontrar patrones de comportamiento.

Una inspección más detallada sobre un conjunto de trabajos y técnicas asociados a la identificación de *outliers* es presentada en la sección 3.7.

3.3. MINERÍA DEL TEXTO

Producto de la facilidad actual de almacenar información, se ha producido un aumento en la cantidad de documentos en lenguaje natural que se encuentran disponibles en formato digital. Sin embargo, debido a que nuestra capacidad para absorber y procesar esta información es limitada, es necesario contar con nuevas herramientas que permitan enfrentar esta problemática.

La minería de texto se define como el proceso que permite descubrir patrones interesantes y nuevos conocimientos en una colección de textos, es decir, se encarga de revelar conocimiento que no existía explícitamente en la colección de textos, pero que surgen luego de relacionar el contenido de varios de ellos. Un aspecto importante de un proceso de la minería de texto es que se debe establecer un equilibrio entre el análisis automatizado y el humano, para que se cumplan los objetivos.

La minería de texto se apoya en un conjunto de técnicas que provienen de la minería de datos, aprendizaje automatizado [55], procesamiento de lenguaje natural, recuperación de información (IR), extracción de información (IE) y administración del conocimiento. Por sus características interdisciplinarias se hacen complicado el abordar todas las áreas en profundidad; por esta razón, en el presente capítulo se cubrirán en detalle sólo aquellas que se consideran relevantes desde el punto de vista del aporte que pueden ofrecer al desarrollo del presente trabajo..

3.3.1. Etapas de la Minería de Texto

En un proceso de Minería de Texto se tienen que realizar un conjunto de etapas para permitir el tratamiento de los datos, estas pueden variar en complejidad, siendo una de las representaciones más básicas la que muestra la Figura 7 obtenida de [17]. Si bien es cierto, este proceso puede ser complementado con un conjunto de recursos lingüísticos adicionales, para efectos prácticos detallar las etapas de la Figura 7 resulta suficiente para obtener una idea sobre el flujo presente en un proceso de minería de texto.

Figura 7: Etapas básicas de minería de texto.

Pre-procesamiento

En minería de texto la etapa de pre-procesamiento se caracteriza por utilizar técnicas de procesamiento de lenguaje natural junto a análisis estadístico y aprendizaje automatizado. Dentro de las tareas que se contemplan en el pre-procesamiento se pueden destacar las siguientes:

- **Limpieza de datos**. Se encarga de eliminar el ruido presente en los datos con el fin de obtener una mejor calidad de estos, lo que afecta directamente a la reducción de la dimensionalidad. Dentro de los alcances de esta tarea se pueden destacar: eliminación de stopwords, corrección de errores ortográficos, normalización de letras (case folding), abreviaciones, eliminación o incorporación de acentos, etc.

- **Representación de documentos.** Los modelos utilizados para la representación de documentos pueden variar en complejidad desde aquellos donde no interesa el orden o las relaciones entre las palabras (bolsa de palabras) a modelos más complejos de tipo n-dimensionales donde se mapean las características de los textos como, por ejemplo vectores, para representar las palabras y sus semejanzas. Estos modelos reciben el nombre de modelos de espacio vectorial.

- **Normalización.** La normalización persigue utilizar una representación estándar para las palabras. En lo que respecta al aspecto morfológico de un lenguaje, los conceptos puede ser formulados de diferentes maneras, lo que recibe el nombre de *polisemia*. Por lo tanto, resulta necesario aplicar transformaciones sobre las palabras, para disminuir la dimensionalidad de términos que comparten una misma raíz. Las transformación corresponde al uso de algoritmos *stemming o lematización*, donde para el caso de *stemming* las palabras se reducen a su raíz *o stem,* mediante la eliminación de los prefijos o sufijos, obteniendo un elemento básico que tendría que mantener las características semánticas del concepto. Entre los algoritmos clásicos de stemming se puede destacar el algoritmo de Porter [21] y el algoritmo de Lovins [22]. Por otro lado la *lematización* se preocupa de encontrar la palabra raíz que corresponde a la forma canónica de un *lexema*.

- **Análisis semántico.** El objetivo principal es tratar la ambigüedad semántica (WSD, por su sigla e inglés). Por otro lado, otra técnica que se encuentra asociada a este tipo de tareas corresponde a la indexación de la

semántica latente (LSI) la cual se utiliza para disminuir la dimensionalidad en aquellos conjuntos de datos muy extensos.

Es importante destacar que en la etapa de pre-procesamiento, están contempladas las tareas de disminución de dimensionalidad y extracción o selección de características, pero estos temas son tratados con un mayor nivel de detalle en la sección 3.5.

En [17] también se puede encontrar una taxonomía de técnicas que están vinculadas a la etapa de pre-procesamiento, la cual se muestra en la Figura 8.

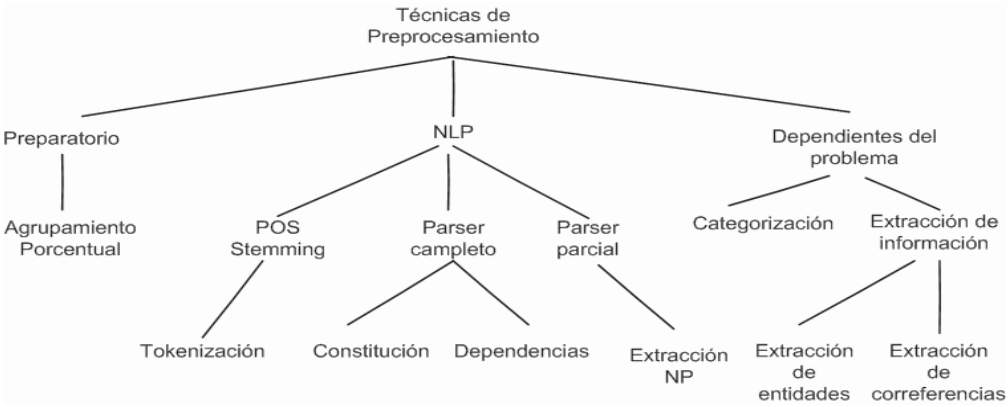

Figura 8: Taxonomía para técnicas de pre-procesamiento.

Operaciones de minería y presentación

En lo que respecta a minería de texto se utilizan tanto técnicas descriptivas como predictivas, para ser empleadas en agrupamientos, clasificación y tareas de extracción de información.

- **Agrupamiento:** Contempla la utilización de algoritmos de aprendizaje no supervisados donde cada item se agrupa en base a la utilización de medidas de distancia. Dentro del contexto de minería de texto, los agrupamientos son utilizados para identificar similitudes entre documentos o palabras. Las técnicas de agrupamiento han sido frecuentemente estudiadas en campos como la recuperación de información, donde uno de los algoritmos mencionados recurrentemente corresponde a K-means.

- **Clasificación:** Las técnicas de clasificación tradicionalmente han estado asociadas a algoritmos supervisados pero en trabajos presentes se manifiesta una tendencia a optimizar la clasificación y al mismo tiempo disminuir el costo del proceso por medio de la utilización de técnicas mixtas [73,74]. En lo que respecta a los algoritmos estadísticos tradicionales se puede destacar a K-NN (nearest neigbors) y Naïve Bayes.

- **Extracción de relaciones:** Al utilizar un modelo de bolsa de palabras no se cuenta con información sobre las relaciones que pueden existir entre las palabras que componen el texto, dentro de las cuales se encuentran las características seleccionadas; por esta razón es necesario utilizar un proceso orientado a

extraer las relaciones entre los elementos. Un enfoque recomendado es el indicado por [17] donde se identifican patrones a partir del análisis de coocurrencias. El análisis está compuesto por dos etapas: (1) la extracción de un conjunto de características candidatas a partir de un análisis de frecuencia de los elementos que aparecen juntos, y (2) la selección de aquellas características presentes en el conjunto obtenido en el paso anterior, que satisfagan alguna medida de calidad.

- **Extracción de entidades:** Contempla la asignación de una etiqueta especial a las entidades (nombres de personas, ciudades, calles, etc.) que ameriten ser recuperadas desde el texto por su importancia semántica. Esta tarea se logra por medio de la etiquetación de la entidad, como resultado de un proceso de clasificación que puede identificar una fuerte relación entre palabras que aparecen junta y el contexto en el cual aparecen.

Visualización

Todo sistema de minería de texto que pretenda potenciar la interacción humana en la exploración y descubrimiento necesita incorporar herramientas gráficas; éstas pueden variar desde recursos muy básicos, donde predominan las herramientas tradicionales, a visualizaciones avanzadas capaces de desplegar conjuntos de datos de alta dimensionalidad que son posibles de explorar en un entorno 3D.

Dentro de las herramientas básicas se puede mencionar a los histogramas (un ejemplo se muestra en la Figura 9). Estas herramientas son de gran utilidad al momento de mostrar los resultados de una consulta, en función de la distribución y proporción de los resultados obtenidos.

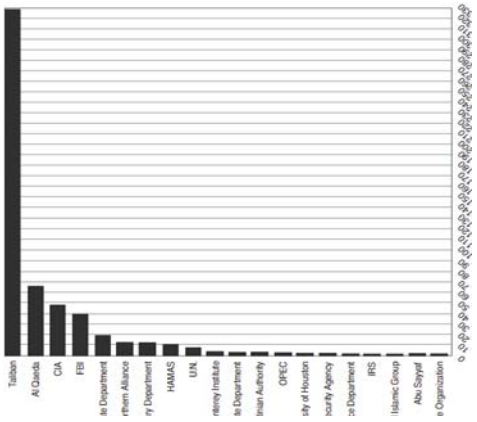

Figura 9: Ejemplo de Histograma [2]

Otra representación que facilita notablemente el trabajo cuando se están buscando reglas de asociación corresponde a la utilización de gráficos circulares. En este tipo de gráficos los elementos son distribuidos alrededor de un anillo y las relaciones se establecen por medio de las conexión entre éstos, donde la "fuerza" de la relación se visualiza por el grosor de las líneas que los vinculan (ver ejemplo en la Figura 10).

[2] Imagen disponible en [17]

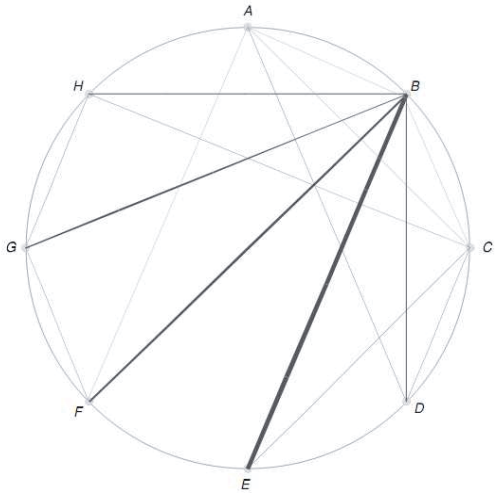

Figura 10: Ejemplo de Gráfico circular

Por último, las herramientas gráficas avanzadas que recurren a la utilización de grafos para ser explorados dinámicamente son frecuentemente utilizadas en los análisis de vinculaciones. Estas herramientas se caracterizan por realizar una representación de red tipo grafo, donde los elementos son los nodos y las relaciones se establecen por los arcos que los vinculan.

Algunas de las ventajas importantes que presentan las herramientas que utilizan estas representaciones gráficas en contraste con las inspecciones orientadas a caracteres son las siguientes:

- **Concisas:** Corresponde a la capacidad de mostrar grandes cantidades de diferentes tipos de datos, todo al mismo tiempo.
- **Relatividad y proximidad:** La capacidad de mostrar agrupamientos, sus tamaños relativos, la similaridad y diferencia de los grupos, y la cantidad de *outliers* que presentan los datos.
- **Focalizar dentro de un contexto:** La posibilidad de manejar una gran cantidad de características mientras éstas son exploradas en un contexto.
- **Acercamiento:** La posibilidad de alternar entre vistas a nivel macro y micro.
- **Estimulación** "rigth brain**":** La posibilidad de involucrar al usuario en un proceso de descubrimiento donde los resultados se obtienen de una manera intuitiva y reactiva, producto de un proceso cognitivo que se lleva a cabo por medio una exploración espacial, que tiene por fin la identificación de patrones.

Este tipo de herramientas cuya potencia ha sido descrita en los puntos anteriores ha adquirido una gran relevancia en estos últimos años donde se aplican en procesos de análisis de datos que reciben el nombre de *minería gráfica.*

Las características que permiten explorar grandes volúmenes de datos por medio de inspecciones visuales sobre grafos 2D o 3D las han proyectado como herramientas idóneas para el análisis de comportamientos en redes sociales, donde es posible aplicar un conjunto de algoritmos que modifican las topologías de la red para facilitar la inspección visual. Adicional a lo anterior, es posible realizar segmentación de redes, agrupamientos de nodos, visualización de atributos como frecuencias, identificadores de distinto tipo, etc.

Pero la utilización en el análisis de redes sociales es sólo una de las áreas de aplicación, estas herramientas también se ocupan con éxito en análisis de seguridad informática. Un ejemplo de esto es la representación de redes de computadores por medio de la estructura de grafos, donde como parte del modelo figuran las comunicaciones que se realizan entre los nodos, lo que permite obtener una visualización de los ataques que recibe una plataforma de comunicaciones con el fin de identificar patrones [95].

Por último, la flexibilidad de esta técnica queda de manifiesto al llevar su aplicación al área del análisis semántico, donde es posible modelar el contenido de textos por medio de una red semántica, la cual tiene por objetivo identificar aquellos conceptos implícitos en los documentos, para determinar el contenido de estos por medio de la observación de las estructuras semánticas que afloran producto de la inspección

La Figura 11 demuestra lo planteado en el párrafo anterior, donde por medio de un caso real se define una red semántica que permite identificar patrones a partir de un análisis de vinculaciones. Dicha red es parte de un trabajo expuesto en FOSDEM 2011 [92], donde la información representada en la red, es obtenida a partir de los documentos diplomáticos de Estados Unidos, liberados por Wikileaks el año 2010.

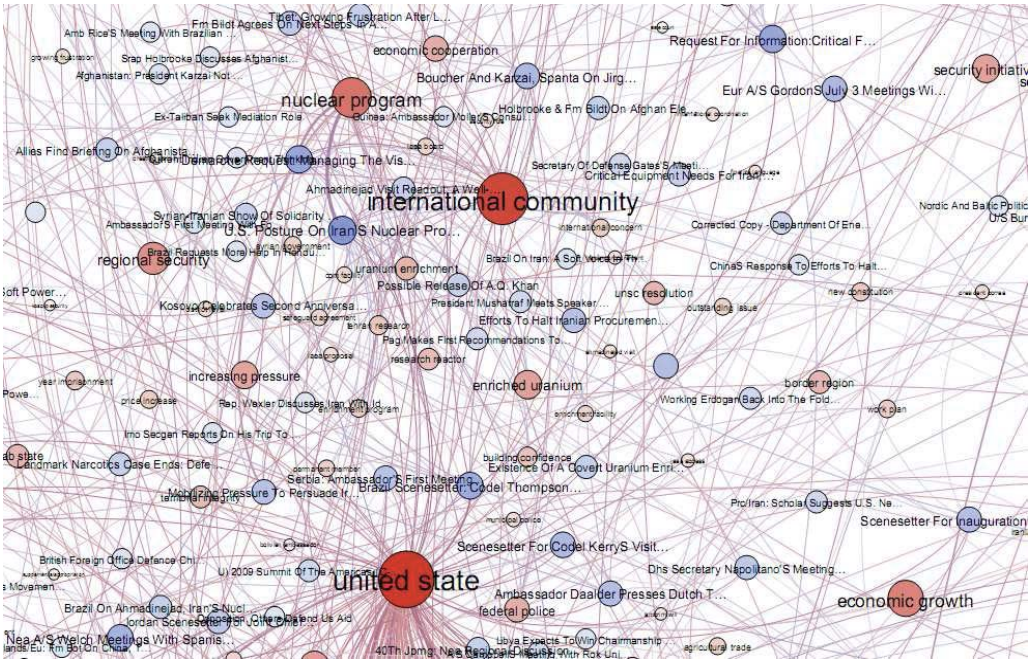

Figura 11: Red semántica para filtraciones de cables, Wikileaks

3.3.2. Recuperación de información

El término Recuperación de Información (IR, Information Retrieval) fue acuñado por Calvin Mooers (1948-1950) para definir un campo interdisciplinario, en donde algunas de las áreas más sobresalientes en las cuales se apoya son: procesamiento de lenguaje natural, informática, álgebra, lógica, matemática, inteligencia artificial, bibliotecología y estadística.

En la literatura asociada, se puede encontrar una enorme cantidad de definiciones para IR, las cuales van variando en el tiempo a medida que se agregan nuevas funcionalidades y técnicas. Por lo anterior, resulta ser de mayor ayuda contar con una definición concisa, que contenga los conceptos fundamentales y se ajuste al sentido tradicional de la recuperación de información, como la que a continuación se presenta:

> "*La recuperación de información busca materiales (usualmente documentos) de naturaleza no estructurada (usualmente texto) desde grandes colecciones para satisfacer las necesidades de información (usualmente servidores locales o Internet*)" [23].

Los sistemas automatizados de IR fueron utilizados en sus orígenes para gestionar grandes volúmenes de información científica (texto no estructurado), desde donde migraron a la administración de los contenidos de grandes bibliotecas, transformándose en una herramienta que sólo podía ser utilizada por un conjunto reducido de especialistas. Hoy el escenario ha cambiado, producto de la actual explosión de contenidos en Internet y en las organizaciones, lo que ha posicionado a los sistemas de IR como una de las herramientas más utilizadas en Internet para la búsqueda de información, la cual no se limita exclusivamente a texto, ya que ahora pueden realizar operaciones de búsqueda en texto semiestructurado, imágenes, videos y archivos de sonido.

En lo relacionado a IR existen algunos aspectos que son considerados relevantes para ser utilizados en el análisis de los datos. Por esta razón se profundiza en la obtención de frecuencias para la obtención de pesos en los términos y en la utilización de modelos para representar documento. Los contenidos que a continuación se presentan son tratados en profundidad en [23, 25, 26, 27,28 ,29].

3.3.2.1. Frecuencia y Peso

Antes de analizar los modelos, resulta apropiado disponer de un mecanismo que permita determinar el nivel de relevancia de los términos que componen un documento. En este punto es donde surge el concepto de frecuencia y *peso*, los que se aplican a los términos que formaran parte de un diccionario, el cual se utiliza para mejorar la relevancia del conjunto de resultados.

Si se realiza una búsqueda en Internet, generalmente se utiliza un conjunto de palabras como patrón de búsqueda sin operadores lógicos (Free Text Query), las cuales son ingresados como una simple cadena de elementos consecutivos que serán utilizados por el motor de búsqueda para seleccionar los documentos que satisfagan la consulta.

En un proceso de búsqueda, si no existe un mecanismo que permita diferenciar las palabras que componen un documento, todas tendrán el mismo nivel de relevancia en el momento de responder la consulta, lo que se puede traducir en la entrega de un conjunto de resultados donde una gran parte no se ajustará al dominio de interés. Una solución a este problema, es asignar un peso en base a la frecuencia de aparición de cada término dentro del documento; de esta forma, el proceso de búsqueda entregará resultados más precisos. Es importante destacar, que dentro del cálculo de frecuencia no se consideran aquellas palabras que no aportan valor a la búsqueda, es decir las *stopwords*.

Formalmente se tiene que el peso corresponderá al número de ocurrencias del término *t* dentro del documento *d*, lo que recibe el nombre de frecuencia del término denotaremos por $tf_{t,d}$. De manera recíproca, se tiene la frecuencia *del documento,* denotado por df_t corresponderá al número de documentos *d* que contienen el término *t*.

Además, si se cuenta con un corpus que contiene **N** documentos, es posible calcular la frecuencia inversa *de documentos* para el término *t* de la siguiente forma:

$$idf_t = \log \frac{\mathrm{N}}{\mathrm{df_t}}$$

El logaritmo presente en la fórmula anterior se encarga del suavizado de los valores cuando se trabaja con colecciones de gran tamaño.

Ahora que se tienen todos los elementos fundamentales, es posible definir el esquema de pesos $tf - idf_{t,d}$, el cual permite asignar al término *t* un peso en el documento *d*, definido por:

$$tf - idf_{t,d} = tf_{t,d} \times idf_t \ = tf_{t,d} \times \log \frac{N}{df_t}$$

Luego, es posible realizar una consulta en base a la utilización de pesos, la cual retornará un conjunto de documentos, que se encuentren ordenados por el grado de similitud que estos presenten, según los términos ingresados en la búsqueda. Este nivel de similitud estará dado por un score obtenido entre la consulta *q* y el documento *d*, el que se genera sumando todos los pesos obtenidos para cada término *t*, correspondientes a la consulta, que se encuentre contenido en el documento *d*.

$$Score(q,d) = \sum t \qquad \text{con } t \in q$$

Es importante aclarar que el calculo de frecuencia y la asignación de pesos no es sólo de utilidad para construir índices, también es de uso frecuente para seleccionar atributos sobresalientes de un texto, con el fin de disminuir la dimensionalidad de los datos, para así focalizarse sobre un conjunto reducido de éstos. Tarea que es común a la recuperación de información y al pre-procesamiento en minería de texto.

3.3.2.2.Modelos para IR

En el campo de la IR existen distintos modelos orientados a representar los documentos y las consultas, con el fin de poder efectuar la recuperación de información en base a estas representaciones.

Antes de profundizar sobre los modelos utilizados en IR, resulta necesario obtener una definición formal que contemple las representaciones de los documentos, las consultas y las comparaciones entre estos elementos. Una explicación propicia para este efecto es la expuesta en [25], donde se define lo siguiente:

Un modelo de IR se define como una cuadrupla compuesta por $[D, Q, \mathrm{F}, R(q_i, d_j)]$, donde:

- **D** representa al conjunto de las representaciones de los documentos de una colección.
- **Q** corresponde a las consultas de los usuarios.
- **F** es el *framework* que permite realizar el modelamiento de los documentos, las consultas y las relaciones entre ambas.
- $\mathbf{R(q_i, d_j)}$ representa una función de puntuación (*ranking*), que asocia valores reales a las consultas $q_i \in \mathbf{Q}$ y al modelado de documentos $d_j \in \mathbf{D}$. Esta puntación permite un orden de recuperación entre los documentos de la colección que se ajustan a la consulta.

Dentro de los modelos más simples y ampliamente utilizados se encuentran los que a continuación se detallan.

Bolsa de palabras

Si se parte de la base que todo documento se encuentra constituido por palabras, se puede decir que el contenido de éste se encuentra estrechamente relacionado con sus elementos constitutivos. Esto tiene relación con el *principio de* composicionalidad*,* el cual indica que la semántica de un documento reside únicamente en los términos que lo forman. Es decir, si una palabra aparece en un documento, éste tratará temas asociados a ésta.

A pesar de la simplicidad y rendimiento apropiados que puede tener este enfoque, en la práctica presenta limitaciones inherentes al principio por el cual se rige, ya que esta representación no considera el orden relativo de los términos lo que no permite determinar el contexto semántico de algunas palabras. Por ejemplo, en la frase “Pedro es más alto que Juan” el orden de las palabras resulta ser importante, ya que no es equivalente obtener un documento del tipo “Juan es más alto que Pedro”.

A pesar de estas limitaciones, este paradigma se ha mantenido vigente por décadas, por su simplicidad y apropiado desempeño.

Modelo Booleano

En este modelo, se extiende la búsqueda a un conjunto de términos, de manera más extendida que en el caso anterior, donde la búsqueda era del tipo *Single Word*. El Modelo Booleano se caracteriza por su simplicidad, ya que utiliza la teoría de conjuntos y el álgebra booleana para definir el patrón de búsqueda. Es decir, para realizar una consulta, el usuario ingresa el requerimiento utilizando una expresión donde los operadores son booleanos (and, or, not) se encargan de concatenar los componentes del patrón de búsqueda. Este tipo de consultas es análogo al realizado sobre las bases de datos que utilizan el leguaje SQL. Del mismo modo, el conjunto de documentos totales se dividirá en dos conjuntos, aquellos que cumplen el criterio de selección y los que no, sin existir algún orden de relevancia sobre los documentos seleccionados. De esta forma, un documento puede ser relevante o no, sin existir estados intermedios. El éxito de este enfoque se basa en su simplicidad, formalidad y su similitud con las consultas de bases de datos.

Sin embargo, este modelo no está libre de limitaciones. Su naturaleza de carácter binario no permite contemplar resultados con un grado de dependencia parcial o diferenciar niveles de relevancia, asociados a los elementos del conjunto de resultados obtenidos. Por lo anterior, todos los términos de la consulta tendrán asignada la misma relevancia, lo que se contrapone a la semántica de un documento, donde los términos presentan distintos niveles de relevancia, los que se pueden asociar al grado de frecuencia presentado en el texto.

3.3.3. Extracción de Información

La extracción de información (IE, Information Extraction), a diferencia de IR, se preocupa de extraer contenidos de interés a partir de fuentes de datos, que por lo general, se basan en un texto no estructurado. Una de las definiciones que podemos encontrar al consultar la literatura asociada es la siguiente:

> *"La extracción de información es la identificación, y la consecuente y concurrente clasificación y estructuración en clases semánticas, de información específica encontrada en fuentes de datos no estructurados, como texto en lenguaje natural, generando información más adecuada para las tareas de procesamiento de información"* [30].

Como el proceso de IE que interesa es el automatizado, se puede agregar que IE se preocupa de la identificación, clasificación, recuperación y estructuración (texto estructurado) de contenidos semánticos presentes en fuentes de texto escritas en lenguaje natural, por medio de la utilización de métodos automatizados. La mayoría de éstos, centrados en técnicas de PLN donde se destacan aquellas basadas en principios estadísticos y de aprendizaje automatizado.

La extracción de información es un componente fundamental en los procesos de minería de texto y últimamente ha adquirido relevancia su utilización para potenciar procesos de IR.

3.3.3.1. El Proceso de Extracción de Información

Un proceso básico de IE cuenta, por lo menos, con cuatro etapas características, dentro de las cuales se ejecutan algoritmos especializados. La Figura 12 presenta las cuatro etapas principales de este proceso, y asociadas a cada una de éstas, un conjunto de técnicas de uso muy frecuentes en las tareas de procesamiento de texto [30].

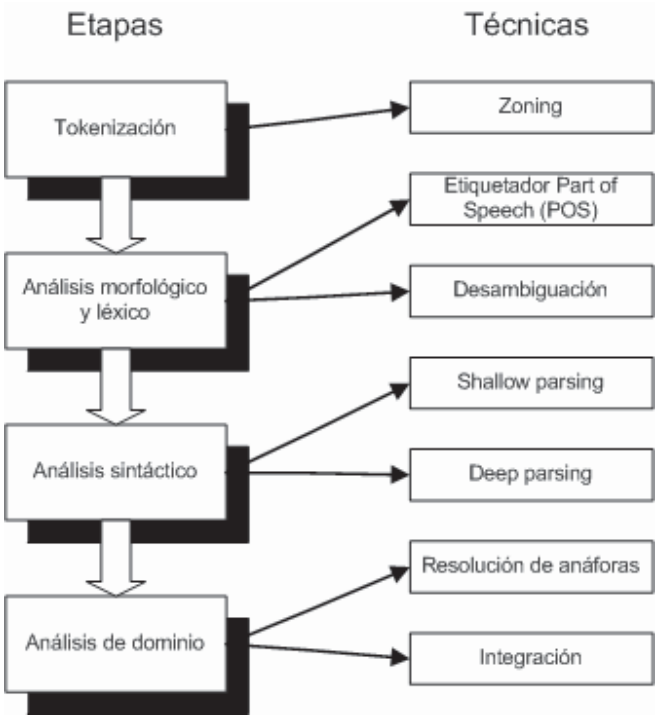

Figura 12: Etapas básicas de Extracción de Información (IE).

Resulta importante aclarar, que tanto las etapas como las tareas que se ejecutan en cada una de éstas pueden cambiar. No necesariamente se tienen que mantener todas las tareas presentes en cada etapa, ya que pueden disminuir o aumentar. Sin embargo, los componentes presentes en la figura anterior suelen ser los que se utilizan con mayor frecuencia.

Tokenización

Se encarga de dividir un archivo de entrada o un string en unidades básicas. Estos bloques por lo general son palabras, sentencias y párrafos.

Análisis morfológico y léxico

La segunda etapa se encarga de trabajar sobre el análisis morfológico y léxico. Para esto, utiliza parser del tipo POS (Part Of Speech) que se encargan de asignar a las palabras una etiqueta que representa la categoría más apropiada

dentro de un contexto. De este modo, se logra la desambiguación de las palabras y se cuenta con información valiosa para ser utilizada en los procesos posteriores.

Análisis sintáctico

En esta etapa se establecen las conexiones entre las diferentes partes de cada sentencia. Esta tarea se realiza con un parser completo (deep o full parsing) o uno superficial (shallow parsing).

Análisis de dominio

En la cuarta y última etapa, se utiliza toda la información provista por los componentes anteriores, para generar *frames* que describen las relaciones entre las entidades.

En los sistemas de IE, al igual que en las tareas que se realizan en PLN, las salidas de algunos algoritmos son utilizados como entrada para otros de mayor nivel, que forman parte de una cadena de proceso, la cual recibe el nombre de pipeline (ver Figura 13). Esta característica también se puede llevar a un nivel macro, donde los componentes ya no son algoritmos si no sistemas especializados de IE, cuyas salidas son utilizadas por otros sistemas, que los preceden en una cadena de procesamiento.

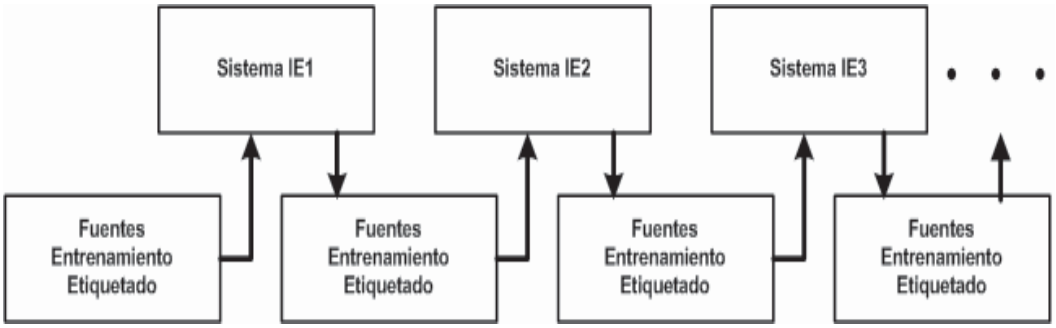

Figura 13: Pipeline entre sistemas de IE.

También es importante aclarar que tanto la IR y IE realizan un pre-procesamiento de texto el cual es muy semejante, donde se busca la transformación del texto a una "forma normal" que facilitará el trabajo de las operaciones posteriores y al mismo tiempo disminuirá la dimensionalidad de los datos.

3.3.4. Distancia de Edición

Introducida por Damerau [31] y Levenshtein [32], define la distancia entre dos strings como la mínima cantidad de ediciones requeridas para convertir uno de ellos en el otro. Por la naturaleza de la forma de calcular esta distancia, primero se tiene que definir un conjunto de operaciones disponibles para realizar la edición. Dichas operaciones se muestran en la Tabla 3.

Operación	Descripción
Inserta	Permite insertar un carácter.
Elimina	Permite eliminar un carácter.
Substituye	Substituye un carácter por otro.
Transpone	Intercambia caracteres en un String, un par a la vez.

Tabla 3: Operaciones que contempla la distancia de edición.

La formalización del cálculo de la distancia de edición se define por:

$$D(A,B) = \min_{j}\left[S(j) + I(j) + R(j)\right]$$

La fórmula indica que la función distancia denotada por *D* entre los strings *A* y *B* corresponde a las mínimas operaciones de substitución *S*, inserción *I* y eliminación *R* que se realizan sobre un string para obtener el otro, dentro de un conjunto de posibilidades definidas por j.

Como la distancia de edición define métricas sobre un conjunto de strings, ésta tiene que cumplir las tres condiciones que constituyen la definición matemática de la distancia, descritas a continuación:

- **Valores no negativos :** $d(a,b) \geq 0; d(a,b) = 0 \text{ Si a} = \text{b}$
- **Simétricas:** $d(a,b) = d(b,a)$
- **Inequidad del triángulo:** Tiene que cumplir el teorema de desigualdad del triángulo

$$d(a,b) + d(a,c) \geq d(a,c)$$

Es importante aclarar, que esta distancia puede ser adaptada de múltiples formas. Un ejemplo de esto, es asignar distintos costos o pesos a las operaciones disponibles con el fin de diferenciarlas.

3.3.5. Expresiones Regulares

Las expresiones regulares (RE, Regular Expressions) corresponden a un lenguaje formal utilizado para realizar búsquedas de strings dentro de un texto. Luego de su creación en la década de los 50's (Kleeme 1956 [33]), esta técnica comenzó a ser utilizada en los entornos Unix. En ese momento, se desarrollaron implementaciones en la mayoría de los lenguajes de alto nivel, herramientas para búsquedas en textos y software de productividad, que son de uso indispensables en tool kits para el procesamiento de lenguaje natural.

En esencia, las RE corresponden a un lenguaje formal que permite realizar búsquedas de strings por medio de notaciones algebraicas. Para lograr su objetivo, las RE requieren de la definición de un patrón, utilizando las notaciones disponibles del lenguaje para que en, base a éste, se puedan realizar las búsquedas dentro de un texto. Como resultado del proceso, se obtendrá un conjunto con todas las ocurrencias que cumplen el patrón de búsqueda.

Patrón de búsqueda

Para definir un patrón de búsqueda se utilizan las notaciones disponibles para crear una RE que permita la identificación de los strings que cumplan con el patrón definido. Es importante destacar que si bien los conceptos son los mismos en las diferentes implementaciones de RE, existen algunas diferencias entre las notaciones y la forma de trabajo que pueden presentar distintos softwares que cuentan con la implementación de un motor de RE. Por esta razón, se aclara que las notaciones que a continuación se presentan son las definidas por la implementación provista por la API NLTK. Se realiza esta referencia explícita debido a que en la parte práctica de esta tesis, la utilización de RE provistas por NLTK, cumplen un papel importante en el pre-procesamiento del texto.

La Tabla 4 entrega un resumen de las notaciones más utilizadas en la definición de RE, las cuales se basan en la implementación provista por NLTK [82]. Una vez conocidos los comandos básicos para realizar las consultas es posible realizar un ejemplo práctico: considerar que se tiene acceso a un corpus compuesto por todos los mensajes electrónicos (un archivo de texto con textos concatenados) recibidos durante un año, y se necesita recuperar todas las direcciones de correo presentes en cada uno de éstos.

Operador	Descripción
.	Comodín, utilizado para representar cualquier caracter.
^abc	Que coincida “abc” al inicio del string
abc$	Que coincida “abc” al final del string
[abc]	Que coincida uno de los caracteres del conjunto (a, b o c).
[A-Z0-9]	Que coincida con un rango de caracteres alfabéticos o numéricos
*	Cero veces o más del ítem anterior; ejemplo: a*
+	Una o más veces del ítem anterior; ejemplo: a+
?	Cero o una vez del ítem anterior
{n}	Exactamente “n” repeticiones, con en entero no negativo.
{n,}	Por lo menos “n” repeticiones
{,n}	No más de “n” repeticiones
{m,n}	Por lo menos “m” y no más de “n” repeticiones
A(b|c)+	Los paréntesis delimitan el ámbito de la operación

Tabla 4: Instrucciones frecuentes para RE.

Teniendo claro el problema, se va a utilizar una ER, la cual tiene por objetivo definir el patrón que identifique a toda dirección de correo electrónico posible de encontrar en el contenedor de correos. Para esto, se define la siguiente RE:

```
[A-Za-z0-9](([_\\.\\-]?[a-zA-Z0-9]+)*)@([A-Za-z0-9]+)(([\\.\\-]?[a-zA-Z0-
9]+)*)\\.([A-Za-z]{2,})
```

El ejemplo permite identificar una cuenta de correo electrónico donde el nombre puede estar formado por letras dígitos o los caracteres “.” y “-“, lo mismo se aplica para la parte de las direcciones del correo que hace referencia al dominio, donde además se definen a lo menos dos niveles (raíz y sub-dominio).

3.3.6. Áreas de Aplicación

La minería de texto se puede aplicar en una infinidad de áreas donde resulta necesario obtener información a partir de texto estructurado, semi-estructurado o no estructurado y para obtener conocimiento a partir de estos. Dentro de las principales áreas de aplicación se pueden destacar las siguientes:

- Motores de búsqueda en Internet.
- Procesamiento de información médica.
- Clasificación de documentos.
- Detección de patrones en correos electrónicos.
- Biología.
- Documentación médica.

Algunas de la áreas mencionadas son tratadas en distintas publicaciones como por ejemplo: lo asociado a detección de fraude es tratado en [41, 19], además existen muchos trabajos vinculados al área médica como por ejemplo [38] y por último en [17] se presenta un listado de áreas de aplicación que pueden complementar aún más lo mencionado en este apartado.

3.4. MINERÍA DE DATOS

La minería de datos es un proceso que se enmarca dentro del área de Descubrimiento de Conocimiento en Base de Datos (Knowledge Discovery in Databases – KDD [59]), el cual a su vez, es parte del proceso de Descubrimiento de conocimiento (Knowledge Discovery- KD). En el ámbito de un proceso KDD, existen múltiples definiciones, para referirse a la minería de datos, dentro de las que destaca la siguiente:

> *"La extracción no trivial de información implícita, previamente desconocida y potencialmente útil a partir de datos"* [59]

Para cumplir los objetivos propuestos, un proceso de Minería de Datos se apoya en áreas como estadística, inteligencia artificial (IA), computación gráfica, bases de datos y el procesamiento masivo de información [34,35].

3.4.1. Etapas de Minería de Datos

El proceso de Minería de Datos, enmarcado en las etapas consideradas en la metodología CRISP-DM [60], contempla las siguientes etapas:

- **Determinación de los objetivos:** Lo primero que se tiene que realizar es determinar los objetivos del cliente, para lograrlo, es necesario que estos sean identificados con la orientación de un especialista en minería de datos. Este proceso puede necesitar de la comprensión del negocio para mejorar el proceso de levantamiento de requerimientos.

- **Pre-procesamiento de los datos:** Comprende la selección, limpieza, el enriquecimiento, la reducción y transformación de los datos. Esta etapa es una de las más costosas en cuanto a consumo de recursos ya que se estima que consume alrededor de un 70% del tiempo total de un proyecto de Minería de Datos.
- **Determinación del modelo:** Para determinar el modelo, se comienza realizando un análisis estadístico de los datos y después se lleva a cabo una visualización gráfica de estos para así, obtener una primera aproximación. En base a los objetivos planteados y las tareas que se deben realizar, se pueden seleccionar un conjunto de algoritmos disponibles, los cuales están vinculados a distintas áreas de la inteligencia artificial. Gran parte de estos algoritmos se encuentran mencionados en la taxonomía presentada en la sección 3.2.
- **Análisis de los resultados:** Corresponde a la última etapa, donde se verifica si los resultados son coherentes y se cotejan con los resultados obtenidos por los análisis estadísticos y de visualización gráfica. El cliente será el encargado de determinar si los datos son novedosos y si estos le aportan algún nuevo conocimiento que le sirva de ayuda para la toma de decisiones.

3.4.2. Áreas de Aplicación

Las áreas de aplicación de técnicas de Minería de Datos pueden ser muy variadas. Sin embargo, las más conocidas son las siguientes:

- Economía.
- En el área médica.
- Detección de fraudes.
- Comportamiento en Internet.
- Meteorología
- Inteligencia

Esta lista sólo representa algunas de las áreas de aplicación que son mencionadas en [37]. Además, en [36] se presenta un conjunto mayor de casos, donde se analiza detalladamente cada uno de ellos.

3.5. REDUCCIÓN DE DIMENSIONES

La reducción de la dimensionalidad, es una de las tareas importantes antes de cualquier proceso de minería de datos o texto, y se aplica cuando el conjunto de variables que se tiene que manejar es demasiado extenso, lo que obliga a disminuir la cantidad de dichas variables para mejorar los resultados que se pretenden obtener. La importancia de aplicar técnicas de reducción en la dimensionalidad de variables según [89] radica en:

- La identificación de un conjunto reducido de características importantes para la predicción de resultados, pueden ser muy útiles desde la perspectiva del descubrimiento de conocimiento.

- Para muchos algoritmos de aprendizaje, la etapa de entrenamiento y el tiempo de clasificación aumenta directamente con la cantidad de características.

- La existencia de ruido o de características irrelevantes puede afectar negativamente a la precisión de un proceso de clasificación.

Para obtener una reducción apropiada existen un conjunto de técnicas disponibles, dentro de las cuales en este trabajo se destacan las asociadas a la minería de texto, ya que son las que se utilizaran en la etapa de diseño.

Por las características propias de los lenguajes (polimorfismo y polisemia), todo proceso de minería que se pretenda utilizar sobre texto tiene que pasar por una etapa de pre-procesamiento orientado a eliminar aquellas características irrelevantes, con el fin de obtener un subconjunto de datos depurados para ser utilizado en las tareas de minería. Dicho conjunto resultante puede ser obtenido por un proceso de selección de características (features selection), el cual se aplica después de realizar el pre-procesamiento de los datos.

A continuación se detallan aquellas tareas que forman parte de la pre-procesamiento, las cuales son utilizadas para disminuir la dimensión de las características de un texto.

- **Pre-procesamiento lingüístico.** Durante la etapa de pre-procesamiento es común utilizar la eliminación de las stopwords, las palabras de tamaño pequeño y la reducción del universo de palabras, por medio de la aplicación de técnicas de stemming y *lematización*. Parte de estas tareas fueron anteriormente mencionadas en detalle en la sección asociada a minería de texto.

- **Categorización semántica.** Durante esta tarea se pretende identificar aquellas palabras o frases que aparecen frecuentemente en el texto, y que además se encuentran directamente asociadas a la semántica del texto (key words). La identificación de n-gramas puede ser de gran utilidad en esta etapa, cuando se pretende encontrar palabras que aparecen juntas dentro de un contexto semántico. Además, producto de esta evaluación se obtiene el vocabulario asociado al texto, lo que se relaciona directamente con la creación de diccionarios de contexto.

- **Reducción estadística.** En este casos se utilizan técnicas estadísticas y de minería para lograr una reducción de la dimensión. Dentro de las técnicas más sobresalientes en esta categoría se pueden destacar la semántica latente (LSI por su sigla en ingles)

Sobre lo antes mencionado, una introducción a nivel conceptual se puede obtener de [17], por otro lado, algunos trabajos que resumen la utilización de técnicas de reducción son los mencionados en [89, 90, 91].

3.6. SELECCIÓN DE LAS CARACTERÍSTICAS

Para la selección de características existen una infinidad de métodos propuestos, los cuales pueden variar en función del área de aplicación, los enfoques aplicados y las herramientas utilizadas, dentro de las cuales se destacan las del área de inteligencia artificial. Algunos de estos métodos están muy ligados a la reducción de dimensiones, tal como fue señalado en la sección anterior.

Pero además de los enfoques anteriores, existen otras alternativas interesantes para la selección de características, tales como las que utilizan modelos de puntuación basados en grafos, como las puntuaciones HINT [93] y el conocido algoritmo PageRanking de Google [94], el cual también se encuentra considerado dentro de los 10 mejores algoritmos para minería de datos [83].

En el caso del algoritmo page ranking, se trata a la Web como un grafo dirigido $G = (V, E)$, donde V representa al conjunto de vértices o nodos y E al conjunto de las relaciones o arcos, que corresponde a un subconjunto de VxV. Para cada V_i existe un conjunto $In(V_i)$, que representa a aquellos nodos que lo apuntan (relaciones de entrada a nodos que lo anteceden), y de manera reciproca existe un conjunto $Out(V_i)$ que representa a los nodos a los cuales apunta V_i (relaciones de salida de V_i a nodos que lo suceden). Utilizando parte de los conceptos anteriores, se obtiene la formula del algoritmo *PageRanking* donde $P(i)$ representa la puntuación de la página i-esima, $P(j)$ es la puntuación de la página que está referenciado a la página a la que se le está calculando la puntuación y $O(j)$ que es igual a $Out(V_i)$, representa las vinculaciones que se establecen a la página i-esima a partir de la j-esima. Por último d es considerado un factor de "amortiguación" cuyo valor varía entre 0 y 1.

$$P(i) = (1-d) + d \sum_{j,i \in E} \frac{P(j)}{O(j)}$$

Además, es necesario agregar que para realizar el cálculo de la puntuación se utiliza la técnica matemática power iteration [96].

Entonces es posible aplicar este algoritmo para la selección de características por medio de la asignación de pesos a los nodos de una red, que esta vez no representa a las páginas Web si no a la estructura semántica de un texto, donde los nodos son las palabras y las relaciones o afluencias entre éstas se representan por el peso de los arcos o aristas, tal como se describe en [97].

De esta forma, se pueden recuperar aquellas palabras que representen de mejor manera los contenidos presentes en el texto, en base a una selección de las mejores puntuaciones que presentan las palabras. Es más, utilizando esta técnica se puede ampliar la recuperación de un nivel de palabras a frases, donde se extraen las palabras que presentan una mayor coocurrencia, recuperándolas como un sólo bloque que contiene una estructura representativa del texto. Por último, se puede agregar que al trabajar con extracciones de conjuntos de palabras, que se encuentran fuertemente relacionadas, la recuperación se puede llevar al área de extracción de entidades, donde las frases recuperadas corresponden a una entidad.

3.7. DETECCIÓN DEL FRAUDE

En lo que a detección de fraude se refiere, el problema en si es complejo y de constante evolución. En un comienzo los mecanismos más utilizados fueron los basados en reglas, similares a los que actualmente utiliza el SNA. En la actualidad se ha evolucionado a la utilización de técnicas de minería de datos [49,64], donde se utilizan sistemas autónomos para reconocimientos de patrones en base a la información histórica disponible.

Un proceso de detección de fraude puede utilizar técnicas de minería de texto, si necesita recuperar datos de fuentes no estructuradas, para posteriormente realizar algún tratamiento sobre estos. También puede utilizar directamente técnicas de minería de datos en el caso de contar con fuentes estructuradas o puede aplicar la utilización combinadas de ambos tipos de minería. En [49] es posible encontrar un conjunto de casos reales donde se muestra la utilización de estas herramientas para enfrentar casos de fraude de distinta naturaleza.

En lo que respecta a las técnicas de minería de datos utilizadas para la detección de fraude, se pueden dividir en dos grupos: descriptivas y predictivas, tal como se vio anteriormente; sin embargo, fuera de esta clasificación lo más importante es saber en qué situaciones se recomienda utilizar alguna de las técnicas. La Tabla 5 tiene por objetivo realizar una clasificación de las técnicas más utilizadas agrupadas por categorías y al mismo tiempo indica en que tipo de problemas se recomienda su utilización.

Tarea	Meta	Técnica de minería
Encontrar datos inusuales	Detectar registros con valores anormales. Detectar múltiples ocurrencias de valores Detectar relaciones entre registros.	Análisis de anomalías.
Identificar relaciones	Determinar perfiles. Determinar registros duplicados. Detección de registros con referencias de valores anormales. Detectar relaciones indirectas entre registros. Detectar registros con combinaciones de valores anormales.	Análisis de *clusters* Análisis de anomalías Análisis de relaciones Asociaciones
Características generales de fraude	Encontrar reglas. Calificación de transacciones sospechosas. Clasificación	Modelos predictivos

Tabla 5: Resumen de técnicas para la detección de fraude.

Dentro de los trabajos realizados en el área de investigación para la detección de fraude, se pueden destacar los siguientes:

- Zengyou He desarrolló un algoritmo de tipo Greedy que resuelve problemas de optimización para la detección de *outliers* de datos categóricos [65]. Además definió un método para la detección de *outliers* [66].

- Kaustav Das considera redes bayesianas para detectar anomalías en grandes conjuntos de datos categóricos [67].

- Tianming Hu aborda la detección de *outliers* a partir de la identificación de patrones obtenidos a través de

técnicas de agrupamiento [68].

- J. A. Fernández Pierna y otros autores realizan un compendio de las principales técnicas utilizadas para la detección de *outliers* [69].
- Mansur y Noor realizan una investigación detallada que describe técnicas para la detección de *outliers* [63].
- Clifton Phua y sus colaboradores [58], realiza un recorrido por los últimos 10 años para obtener un compendio de trabajos sobresalientes, donde se utiliza la minería de datos para enfrentar la detección de fraudes.
- Por otro lado, un trabajo que define un proceso detallado digno de ser mencionado es [57], donde se realiza el uso conjunto de técnicas de *backpropagation*, Naive Bayes y árboles de decisión (C4.5) para la detección de fraude en casos de seguros automotrices. Además, como una forma introductoria a las herramientas disponibles para la detección de fraude se encuentra la publicación de Jesús Mena [41].

Los trabajos anteriores permiten tener una idea de las distintas técnicas que se suelen utilizar para generar soluciones para la detección de fraudes. Sin embargo, en este trabajo se ha optado por realizar inspecciones en base a algoritmos de agrupamiento no supervisadas, ya que se desconoce el comportamiento de los datos, debido a que se trabaja con precios de productos y estos varían constantemente. La identificación de anomalías por medio de la utilización de agrupamientos permite descubrir a aquellos objetos que presentan una mayor distancia con el resto de la población. De este modo, mientras mayor sea la distancia entre un objeto y el resto de la población, mayor es la posibilidad de considerarlo un *outlier*.

Resumiendo, se tiene que las técnicas de detección de *outliers* se pueden clasificar de la siguiente manera [99]:

- **Basadas en modelos**. Se refugian en la estadística para reconocer la distribución de los datos. Entre estas técnicas se destacan: método de incertidumbre y de "convex hull".
- **Basadas en proximidad**. Se apoyan en el manejo de la distancia entre los objetos; mientras mayor sea la distancia que separa a los objetos de un comportamiento generalizado, éstos pasan a considerarse *outliers*.
- **Basadas en densidad**. Se sustenta en la estimación de la densidad de regiones de objetos. Para esto, los objetos ubicados en regiones de baja densidad se consideran como anómalos. Este método resulta apropiado para detectar anomalías donde las áreas altamente pobladas contienen valores que pueden ser considerados como normales. Un algoritmo utilizado para la detección de *outliers* en base a densidad es LOF, el cual se detalla a continuación.

Local Outlier Factors (LOF)

El algoritmo LOF [100], se basa en la densidad de un conjunto de objetos para asignar a cada uno de estos un grado de atipicidad, cuyo valor recibe el nombre de Local Outlier Factor (LOF). Lleva por nombre "local" ya que el valor del factor depende de la relación que existe entre objetos aislados con el vecindario circundante.

Los pasos que realiza el algoritmo se pueden resumir de la siguiente forma:

1. Por cada punto *O* se calcula *k-distance* (distancia al k-esimo vecino más cercano).
2. Se calcula la distancia de accesibilidad para cada punto *O* con respecto a *P*. , denotado por reach-dist(O.P) = *MAX {k-distance(p), d(o,p)}*, donde *d(o,p)* es la distancia desde *O* al punto, que corresponde a la densidad local de accesibilidad y está dado por el inverso de la distancia media entre *O* y el vecindario. Por otro lado, $MinPts$ (por lo general se utiliza *k*), es el valor que define el vecindario a crear alrededor de O.
3. Se calcula la densidad local de accesibilidad de O ($lrd_{MinPts}(o)$) basado en el valor de $MinPts$.
4. Se calcula el valor de LOF dado por la media entre los coeficientes de la densidad local de accesibilidad de P y los puntos vecinos más cercanos.

La formula para obtener la puntuación LOF es la siguiente:

$$LOF_{MinPts}(p) = \frac{\sum_{o \subset N_{MinPts}(p)} \frac{lrd_{MinPts}(0)}{lrd_{MinPts}(p)}}{\left| N_{MinPts}(p) \right|}$$

3.8. CLASIFICACIÓN DE PRODUCTOS

Asociado a la clasificación de productos, han sido fundamentales los trabajos realizados por Rayid Ghani [74,77,79,81] y Katharina Probst [77,81]. En estos casos la clasificación de los productos, se basa en la identificación de los atributos que los componen, utilizando técnicas de minería de texto para trabajar con la semántica presente en los campos de descripciones. Sin embargo lo más sobresaliente de estos estudios, es la incorporación de co-entrenamiento donde se definen dos vistas: una orientada a realizar una clasificación por medio de un algoritmo supervisado y la otra utilizando un método como ME para expandir el aprendizaje a datos no etiquetados.

En la metodología antes señalada se destaca lo siguiente:

- Un método automatizado que permite clasificar productos con una mínima interacción humana. El sistema aprende tanto de manera supervisada como semi-supervisada, lo que disminuye los costos vinculados a la etiquetación de casos de prueba, ya que los datos etiquetados son mínimos y sólo se utilizan como semilla.
- Los resultados obtenidos por medio de la incorporación de minería de texto, permiten ampliar el ámbito del análisis alcanzado por la minería de datos, profundizando éste sobre aspectos que antes se encontraban fuera de alcance.

Uno de las limitaciones de los análisis de minería de datos en que muchos autores concuerdan, corresponde a la

limitación que ésta presenta para trabajar con la semántica de los datos. La incorporación de técnicas de minería de texto, permite trabajar con dicha semántica, facilitando la inferencia de los atributos semánticos de los productos logrando su individualización, lo que permite realizar análisis más detallados sobre los datos que se están estudiando.

3.9. ATENUACIÓN DE RUIDO EN LOS DATOS

En base a las reuniones realizadas con personal de Aduana, se contaba con el conocimiento previo que indicaba que los datos a analizar presentarían un nivel significativo de distorsiones. Por esta razón, se procede a incorporar dentro del estado del arte, aquellas publicaciones asociadas a la atenuación del ruido en los datos.

Uno de los trabajos más interesantes en este tema se describe en [71], donde se plantea la utilización de un *framework* para realizar la minería de opiniones, destacando las distintas manifestaciones de ruido presentes en los datos y las medidas aplicadas para su atenuación. Existen trabajos adicionales como el de Kernighan [86], que utiliza el paradigma del canal con ruido de Shanon [10] para realizar un corrector ortográfico, el que será optimizado posteriormente por Brill [84, 85].

Otro trabajo importante de destacar es Clark [71], quien presenta una metodología basada en aprendizaje automatizado utilizando modelos generativos junto al paradigma del canal con ruido.

Por último, se puede mencionar el trabajo de Nasukawa [72], quien realiza un interesante estudio centrado en la delimitación de sentencias utilizando las pausas de información para generar un modelo en base a N-gramas.

3.10. CONCLUSIONES

En el estado del arte se inspeccionaron un conjunto de técnicas y trabajos de investigación que se encuentran dentro de los alcances de esta tesis. Las áreas de recuperación y extracción de información han sido de gran importancia ya que en base a la primera se ha podido profundizar en las técnicas de pre-procesamiento de datos y de la segunda se han rescatado los elementos necesarios para extraer entidades a partir de datos no estructurados, ambos elementos fundamentales para el diseño.

Sobre la literatura recolectada, es importante mencionar que no se han encontrado trabajos que cubran directamente el área de interés, o que entreguen una solución orientada a la detección de fraude donde el tratamiento del ruido en los datos sea tratado como un elemento importante dentro de la solución.

Por último, en el capítulo siguiente se retomaran muchos de los elementos mencionados en el estado del arte, los cuales serán adaptados e integrados para transformarse en los elementos constitutivos del *framework.*

4. DISEÑO DEL FRAMEWORK

El presente capítulo tiene por objetivo detallar los productos de software seleccionados para la implementación de la solución, así como también describir cada uno de los componentes utilizados a nivel de cada una de las etapas junto a las funcionalidades de los algoritmos y técnicas empleadas.

Como dato adicional, se aclara que los resultados que se obtienen aplicando el *framework* a la partida arancelaria seleccionada para el análisis, serán expuestos en el capítulo quinto.

4.1. METODOLOGÍA

Considerando que el presente trabajo corresponde a una investigación en la cual participa sólo una persona, por lo tanto no existe la complejidad inherente de administrar distintos grupos de trabajo, y además no se presentan obligaciones que impliquen el cumplimiento de formalidades con la organización dueña de los datos, resulta conveniente optar por un planteamiento metodológico que se caracterice por su simplicidad y flexibilidad, pero que asegure el cumplimiento de los resultados.

Por lo anterior, se ha optado por definir la siguiente metodología, la cual cuenta con el apoyo de las etapas del proceso KDD, en lo que respecta a las tareas orientadas al descubrimiento de conocimiento.

1. **Recopilación de documentación**. Esta etapa contempla la recopilación de toda la información asociada a la organización y a los avances técnicos en las áreas de estudio, la cual posteriormente será utilizada para conformar el estado del arte.

2. **Análisis de datos.** Comprende la inspección de los datos fuentes, para identificar las características de éstos y el nivel de ruido que presentan. En esta etapa se realiza el análisis de frecuencia y de relaciones sobre los datos que fueron seleccionados.

3. **Creación de prototipos**. Comprende la construcción de los componentes que formarán parte del *framework.* Para el caso de minería de texto, estos corresponden a módulos que tienen que ser desarrollados; por el lado de la minería de datos, el producto corresponde a un *workflow* de trabajo diseñado en la herramienta Clementine, para realizar agrupamientos con el fin de identificar *outliers*.

4. **Evaluación de prototipos**. Se evalúa cada componente del *framework* por separado.

5. **Integración**. Se realiza la integración de cada uno de los componentes antes evaluados, para conectar las distintas tareas en pipeline.

6. **Evaluación global**. Con el producto integrado, se realizan pruebas sobre los datos y los resultados obtenidos son evaluados por expertos.

Además de lo antes indicado, se tiene que destacar que los puntos 2, 3 y 4 están basados en un proceso KDD donde cada prototipo es mejorado por medio de iteraciones sucesivas.

4.2. HERRAMIENTAS DE SOFTWARE

Para el diseño y desarrollo del *framework* fue necesario recurrir a un conjunto de herramientas y productos de software, los cuales fueron seleccionados, entre otros aspectos, por características que se consideraron relevantes para facilitar el trabajo asociado a la manipulación y análisis de datos. Los productos seleccionados son mencionados a continuación, junto con las respectivas justificaciones.

a) Base de datos

La información de la fuente de datos utilizada corresponde al período de transacciones del año 2010, que forman parte de la base de datos histórica que utiliza el Servicio Nacional de Aduana. A partir de la estructura de la base de datos, se seleccionaron las tablas que contenían los datos de interés, para posteriormente recuperarlos a través de la utilización del entorno de administración y consultas provisto por la herramienta Toad. Las características de los productos involucrados en el almacenamiento de los datos y la herramienta de administración son indicados a continuación.

Oracle

Ítem	Descripción
Producto	Oracle DB Enterprise Edition
Empresa	Oracle
Versión	10G 10.2
Plataforma OS	Linux Red Hat AS 4.0
Justificación	La versión de la base de datos es la misma que se utiliza en la base histórica. Se optó por emular el entorno para facilitar las tareas de replicación de datos.

Toad

Ítem	Descripción
Producto	Toad for Oracle
Empresa	Quest Software
Versión	8.6.1.0
Plataforma OS	Windows
Justificación	Se utilizó Toad por ser uno de los entornos de desarrollo y administración que posee reconocida madurez y posicionamiento de mercado. Es un entorno que facilita trabajar con la base de datos y al mismo tiempo es una herramienta conocida por el autor de esta tesis.

b) Virtualización

Se utilizó para replicar el entorno de la base de datos histórica, salvo los datos sensibles, con el fin de utilizar sólo información de dominio público, pero en una plataforma portable. El entorno virtualizado utiliza VMWare Player para levantar un OS Red Hat Linux, con una base de datos Oracle.

VMWare Player

Ítem	Descripción
Producto	VMWare Player
Empresa	VMWare
Versión	3.0.0
Plataforma OS	Windows
Justificación	Las principales razones para trabajar con Virtualización son: 1. Portabilidad, los datos pueden ser transportados con su entorno sin mayores problemas 2. Permite mantener copias de todo el entorno virtual de manera muy simple 3. La virtualización por medio de VMWare, obedece principalmente a que dicho producto se encuentra posicionado como uno de los líderes del área, el cliente es gratuito y existe experiencia previa sobre su utilización. 4. La virtualización permite replicar un proceso de negocio lo más cercano a la realidad, pero bajo un entorno controlado y portable.

c) APIs utilizadas

Para la etapa de minería de texto, fue necesario buscar un producto que entregara una buena cobertura de las herramientas de procesamiento de lenguaje natural, las cuales son fundamentales para el procesamiento de los campos no estructurados. Por otro lado, en lo referente a la visualización fue preciso incorporar una API que a partir de los datos permitiera generar una red de coocurrencias, con el fin de utilizar esta estructura en un visualizador que permita la exploración y manipulación. A continuación se detallan las fichas de cada uno de los productos antes mencionados

NLTK

Ítem	Descripción
Producto	Natural Language ToolKit
Empresa	Google Code
Versión	2.0B
Plataforma OS	Multiplataforma
Justificación	La API fue seleccionada por las siguientes razones: 1. Orientada para el aprendizaje de PLN 2. Soporte multi-lenguaje 3. Bien documentada 4. Se entrega junto con Demos de programación 5. Existen publicaciones sobre aspectos avanzados.

Networkx

Ítem	Descripción
Producto	Networkx
Empresa	Los alamos nacional laboratory
Versión	1.5
Plataforma OS	Multiplataforma
Justificación	Excelente API que permite trabajar con grafos utilizando Python. Además de los métodos que permiten crear y manipular una red, ofrece un conjunto de algoritmo para realizar operaciones científicas sobre ésta.

d) Lenguaje de programación

En lo que respecta a los lenguajes de programación, para la recuperación de las transacciones desde la base de datos histórica se utilizó un *script* escrito en lenguaje de consulta SQL y para generar todos los programas asociados a las etapas comprendidas dentro de la minería de texto se utilizó Python potenciado con las APIs mencionadas en el tópico siguiente.

Python

Ítem	Descripción
Producto	Python
Empresa	Python software foundation
Versión	2.6.5
Plataforma OS	Multiplataforma
Justificación	Python se utiliza principalmente en las etapas del análisis preliminar y el preprocesamiento de datos. Las operaciones que están asociadas al trabajo con Tokens y cálculos de frecuencias se realizan por medio de la API NLTK, la cual se encuentra basada en Python. El lenguaje posee características deseables para el procesamiento de lenguaje natural, dentro de las cuales se pueden destacar: 1. Orientado a objetos. 2. Simple pero poderoso. 3. Instrucciones importantes para el manejo de texto. 4. Su curva de aprendizaje no es elevada. 5. Posee soporte para gráficos, procesamiento numérico, Web, etc. 6. Cuenta con un conjunto de extensiones que permiten ampliar su ámbito de acción a distintas áreas por medio de la incorporación de componentes para el desarrollo de aplicaciones.

SQL

Ítem	Descripción
Producto	SQL
Empresa	Oracle
Versión	10.2.0.1
Plataforma OS	Se encuentra implementado en muchos productos para distintas plataformas de sistemas operativos.
Justificación	El lenguaje SQL se utiliza en la etapa de recuperación de los datos presentes en la base de datos Oracle. La creación del script se realiza por medio de la herramienta Toad, que permite la creación, depuración y ejecución de los scripts SQL. Se utilizó el soporte SQL de Oracle ya que viene nativo con la base de datos que se está utilizando. En este caso la versión del producto que figura en esta ficha técnica corresponde a la versión del motor SQL integrado en la base de datos Oracle

e) Entorno de desarrollo

Para la selección del entorno de desarrollo primó el conocimiento previo que el autor tiene sobre el IDE Eclipse, el cual por su capacidad de utilizar *plugins* puede transformarse en una plataforma de desarrollo para múltiples lenguajes, entre estos Python. Algunas de las características que han transformado a este producto e uno de los IDEs favoritos de programadores son expuestas en la siguiente ficha técnica

Eclipse

Ítem	Descripción
Producto	Eclipse SDK
Empresa	Eclipse
Versión	3.4.2 Ganymede
Plataforma OS	Windows
Justificación	La selección de este IDE como entorno de desarrollo se justifica por las siguientes razones: 1. Entorno multiplataforma. 2. La integración a distintos lenguajes es simple y se realiza por medio de plugin, manteniendo la estructura base de trabajo que ofrece el IDE. 3. Soporte integrado de desarrollo y debug. 4. Fácil de instalar y de replicar un entorno ya configurado. 5. Soporta control de versiones 6. Integrado con ANT 7. Dispone de asistentes

f) Entorno para Minería de Datos.

En la etapa de minería de datos se necesitaba de una herramienta flexible que permitiera un desarrollo de un flujo de trabajo de manera rápida y gráfica, por medio de la incorporación de distintos componentes para la confección de

prototipos, donde las transformaciones que se realizaban sobre los datos y los resultados obtenidos tenían que ser de fácil comprensión, con el sólo hecho de inspeccionar el flujo de trabajo. Por estas razones, más las indicadas en la ficha técnica, fue seleccionado el producto Clementine.

Clementine

Ítem	Descripción
Producto	Clementine
Empresa	SPSS
Versión	12
Plataforma OS	Windows
Justificación	El entorno de trabajo Clementine para minería de datos fue seleccionado por lo siguiente: 1. Trabajo en workflow que permite una fácil comprensión de los procesos que trabajan en *pipeline.* 2. Completo soporte de herramientas para minería de datos. 3. Producto que se encuentra dentro de los mejores del área. 4. Curva de aprendizaje no pronunciada. 5. Amplio soporte para representaciones gráficas

g) Control de versiones.

Tiene el objetivo de facilitar la administración y sincronización tanto de los documentos que componen la tesis, como los datos de pruebas y el código fuente de los programas. Se decidió instalar un servidor SVN para este propósito, junto al respectivo cliente

SVN Server

Ítem	Descripción
Producto	Subversion
Empresa	Tigris
Versión	1.4.2
Plataforma OS	Centos 5.0
Justificación	Subversión es un sistema de control de versiones mucho más eficiente que CVS, con un conjunto de funcionalidades agregadas para superar las limitaciones de su predecesor. Además su utilización se justifica por el valioso apoyo para la administración de versiones y por tratarse un sistema de control de versiones de código y documentos, que se utiliza bastante en los entornos de desarrollos.

Tortoise SVN

Ítem	Descripción
Producto	Tortoise SVN
Empresa	Tigres
Versión	1.6.7
Plataforma OS	Windows
Justificación	El cliente Tortoise es uno de los más utilizados con subversión, dentro de las comunidades open source. Se caracteriza por ser intuitivo, posee una GUI integrada con el explorador de archivos de Windows, lo que facilita notablemente su utilización.

h) Visualización

Para permitir realizar operaciones interactivas que faciliten las tareas de descubrimiento para los analistas, se han incorporado herramientas que permiten una exploración de los datos de manera gráfica. De este modo se pretende incorporar al *framework* la potencialidad de la minería gráfica por medio de la exploración de estructuras de grafos que pueden ser utilizadas para representar redes semánticas, que muestren las relaciones que se dan entre los términos, la frecuencia de estos y al mismo tiempo permita identificar aquellas palabras que con mayor frecuencia aparecen juntas.

Existe una gran cantidad de herramientas pagadas y gratuitas que permiten explorar y crear redes que se representan por medio de estructuras de grafos, pero se decidió ocupar el producto Gephi por las razones que se resumen en la siguiente ficha.

Gephi

Ítem	Descripción
Producto	Gephi
Empresa	Gephi
Versión	0.8 Alpha
Plataforma OS	Multiplataforma
Justificación	Producto open source que a pesar de mantener una versión alpha ha adquirido gran popularidad dentro de las herramientas de visualización y manipulación de redes. Dentro de sus puntos fuertes destacan su amplia compatibilidad con distintos formatos de archivos para representación de redes, una amplia gama de algoritmos para manejar topologías y su excelente gráfica.

4.3. MACRO ETAPAS

En *framework* puede ser resumido en seis macro componentes, donde las etapas de pre-procesamiento y minería de texto contemplan un flujo de retroalimentación, el que se utiliza gracias al apoyo de la etapa de análisis previo y la interacción de ésta con los resultados de los recursos lingüísticos obtenidos por medio de la creación de los diccionarios. A la vez, la totalidad de estas etapas se comunican entre si en un esquema de pipeline, donde la salida obtenida por uno es utilizada como entrada por el siguiente. Las herramientas seleccionadas para la implementación son: NLTK por ser una API para Python que facilita mucho el trabajo con strings, entregando además un conjunto de algoritmos de utilidad para minería de texto y Clementine, seleccionado por presentar un entorno de trabajo intuitivo, basado en un *workflow* que facilita la creación y presentación de prototipos junto a un amplio soporte de algoritmos y herramientas para minería de datos.

La Figura 14, se preocupa de esquematizar la interacción entre los componentes, donde además se especifica el lenguaje y API utilizadas para desarrollar los componentes de minería de texto, así como también el producto que se utiliza para efectuar las tareas de minería de datos.

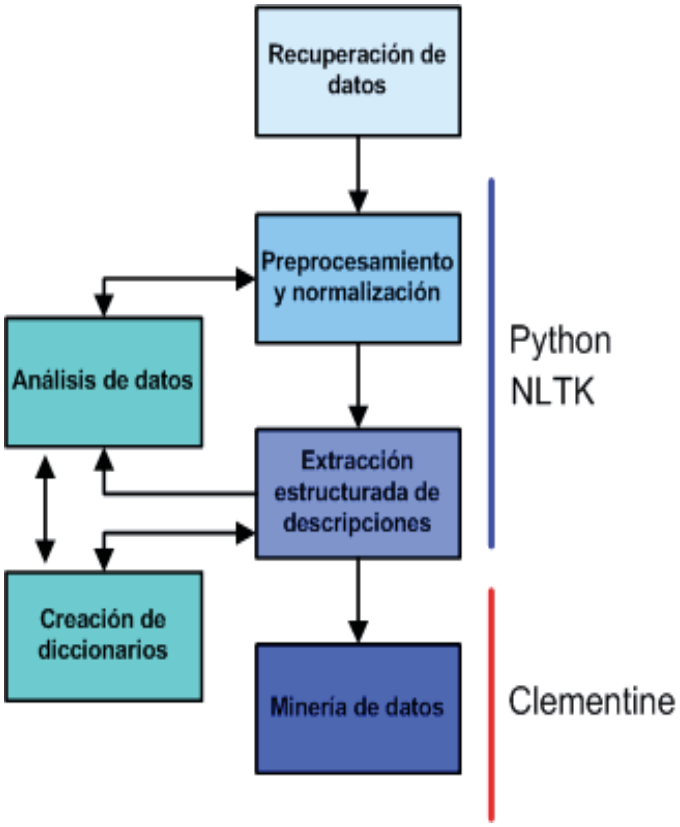

Figura 14: Macro etapas del framework

La representación de la figura anterior, tiene por objetivo entregar una representación macro, la cual a continuación será inspeccionada por medio de un análisis *Top-Down* que permita una revisión ordenada desde lo más general a o más particular, de esta forma se pretende clarificar los procesos que se llevan a cabo en cada una de las etapas y las tareas encargadas de realizar las distintas transformaciones.

4.4. DISEÑO DETALLADO DE LAS ETAPAS

En esta sección se presenta una descripción detalla de cada una de las etapas que componen el *framework*, profundizando en las decisiones de diseño que fueron tomadas y en los resultados que se esperan obtener. Además, en el diseño se detalla el proceso para generar los diccionarios, recursos lingüísticos que cumplen un papel fundamental durante la minería de texto.

4.4.1. Recuperación de Datos

Esta etapa corresponde a la recuperación de los registros que se encuentran en una base de datos históricos que contiene las transacciones realizadas para las importaciones y exportaciones durante el año 2010. De las transacciones de importaciones (DIN) se seleccionan un subconjunto de datos para ser utilizados tanto en el análisis previo como en la selección de los conjuntos de entrenamiento y evaluación.

Para la recuperación de datos se creó un script SQL que se encarga de extraer los registros de la base de datos histórica para la partida arancelaria T-Shirts (codificación arancelaria 62052010 y 61091011). El script genera como salida una planilla Excel que contiene los registros asociados a las transacciones realizadas durante el año 2010 para las importaciones vinculadas a la partida arancelaria de interés. Los campos que se recuperan de los documentos de importaciones se indican en la Tabla 6.

Campo	Descripción
ID	Es el identificador del registro y corresponde al ID de la declaración de importación DIN.
País	Código que identifica al país de origen
Puerto	Puerto de entrada de las mercancías
Aduana	Aduana que recepciona las mercancías
Cantidad	Cantidad de unidades importadas
Valor total	Valor total de la importación para el productos
Valor unitario	Valor unitario de cada producto
Descripción	Descripción del producto en texto no estructurado

Tabla 6: Campos recuperados para las importaciones

El último campo que figura en la Tabla 6, corresponde a la descripción del producto en texto no estructurado. Dicho campo será el centro del estudio en la etapa de minería de texto, ya que sobre éste se tendrán que recuperar los atributos que permitan individualizar a un producto.

La salida de la etapa corresponde a un archivo en formato CSV que contiene los campos de la Tabla 6, el cual es utilizado como entrada en la etapa siguiente.

4.4.2. Pre-procesamiento y Normalización

El pre-procesamiento y el análisis previo son dos etapas que se potencian mutuamente, con el objetivo de mejorar los resultados de salida en base a un proceso de ajuste iterativo. Pero para mantener una coherencia con la secuencialidad del flujo de procesamiento, se ha optado por considerar a ésta como la etapa siguiente a la recuperación de datos.

Primero se define la estructura que se utilizará para trabajar con las transacciones de importaciones. Dicha estructura consolida todas las transacciones junto con la descripción de productos, donde cada registro es representado por una fila en un archivo de texto, de este modo, una fila es una transacción de importación; esto permite disminuir el acceso a disco a diferencia de un enfoque donde cada transacción es un archivo, además se cuenta con una estructura adecuada para trabajar con descripciones, las cuales en promedio contienen 45 tokens. Por cada partida arancelaria que se analiza se obtiene un archivo consolidado con las descripciones de las importaciones obtenidas para el período de estudio. La Figura 15 esquematiza el proceso de consolidación.

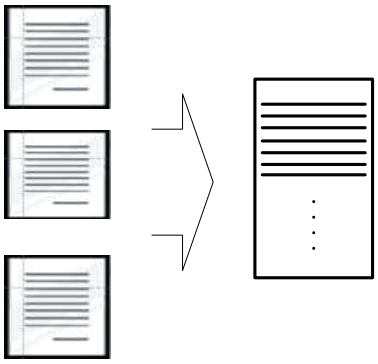

Figura 15: Concatenación de importaciones

Una vez obtenida la estructura de trabajo se procede a realizar el pre-procesamiento del texto, donde lo primero es fragmentar el texto en tokens para luego "normalizarlo", de esta forma es posible disminuir su dimensionalidad y al mismo tiempo se facilita el trabajo a realizar en las etapas posteriores. La etapa de pre-procesamiento es compartida tanto por la etapa de análisis previo como por la de extracción estructurada de descripciones. Dentro de las tareas que se realizan durante el pre-procesamiento y normalización se destacar las siguientes:

1. **Encoding:** La primera tarea que se realiza es la conversión de encoding, donde se codifica el archivo concatenado como ANSI ASCII. Si bien es cierto la codificación no soporta caracteres extendidos, eso no afecta negativamente al proceso ya que la base de datos de donde se recuperan los registros posee las mismas limitaciones.

2. **Tokenización:** Se encarga de segmentar el texto en tokens, donde los signos de puntuación no se consideran tokens. El criterio de segmentación utiliza los espacios en blancos para delimitar las fronteras de los tokens, donde no se realizar recuperación de tokens compuestos por múltiples palabras o elementos.

3. **Safe-Case:** Implica la transformación a minúsculas de todos los tokens que contengan caracteres alfabéticos. Esto disminuye la dimensionalidad de los datos, no afectando al proceso, ya que éste no contempla la utilización de mayúsculas para identificar palabras que representan a entidades o personas.

4. **Eliminación de puntuación:** Todos los tokens que contienen caracteres de puntuación son eliminados. La puntuación no entrega ningún valor al análisis ya que en las descripciones no se cuenta con texto riguroso, en cuanto a la utilización de reglas de puntuación. Por el contrario, el texto de las descripciones no se preocupa de respetar reglas gramaticales, debido a que su contenido es minimalista en base a un lenguaje técnico, muy acotado, donde sólo se pretende enumeran características.

5. **Eliminación de símbolos:** Se elimina toda simbología que no tenga sentido dentro del contexto. Sólo se mantienen aquellos tokens cuya significado está asociado a unidades de medida y de proporción.

6. **Eliminación de paréntesis:** Se comprobó que los distintos tipos de paréntesis no entregan valor alguno a las descripciones, por esta razón son eliminados.

7. **Eliminación de caracteres repetidos:** En los datos se detectó una gran cantidad de caracteres repetidos que aparecen en forma consecutiva en el texto, los cuales se manifiestan como letras o signos de puntuación. Todos estos caracteres tienen que ser eliminadas durante el pre- procesamiento.

8. **Eliminación de espacios:** Como parte de la normalización no se consideran los espacios en blanco adicionales a los que delimitan la frontera entre tokens.

9. **Eliminación de *Stopwords*:** Todo pre-procesamiento de texto contempla la eliminación de las *stopwords*, ya que son palabras extremadamente recurrentes en el texto, lo que las transforma en poco útiles para los procesos de clasificación. Por tratarse de texto en español este proceso utiliza el conjunto de stopwords para el español que entrega NLTK el que fue complementado con otras fuentes para lograr un listado de palabras lo más completo posible.

10. **Eliminación de inflexiones:** Se identificó que una de las causas importantes que aumentaba la dimensionalidad de las palabras estaba asociado a la pluralización de algunos términos. Para solucionar esto se desarrolló un componente para el manejo de inflexiones que se encarga de singularizar cada token.

11. **Aplicación de sinónimos:** Como parte del texto existen muchas palabras en ingles las cuales son tratadas como sinónimos con el fin de encontrar su equivalente a español, de esta forma se pretende eliminar toda referencia de palabras en ingles del texto. Para efectuar esta tarea se utiliza un diccionario de sinónimos que se encarga de las conversiones.

4.4.3. Análisis de Datos

Antes de realizar cualquier proceso de clasificación sobre los datos, es necesario primero tener una idea de las características de los elementos (tokens) que componían las descripciones de los productos. Así se puede contar con una visión preliminar sobre la posible complejidad que se tiene que enfrentar en el momento de realizar el procesamiento de éstos.

Para el análisis de los textos se utiliza un modelos de bolsa de palabras, enfoque para el cual se trabajó con un conjunto de herramientas que se encargan de entregar una visión sobre las características de los datos, y al mismo tiempo apoyan a la tarea de selección de características, las que serán utilizadas en las tareas de clasificación. Las herramientas desarrolladas cumplen con implementar las tareas iniciales de un proceso NLP, donde el texto se transforma en tokens, para luego realizar un análisis sobre éstos, con la finalidad de identificar los tipos de datos, las frecuencias, las relaciones que se establecen entre los datos, para terminar con una selección de características.

Dentro de las salidas importantes que generan las distintas herramientas se encuentran las puntuaciones de tokens por tipos, por frecuencias y visualización de las coocurrencias por medio de la obtención de una red semántica, que representa gráficamente las relaciones y las frecuencias de los tokens. Las visualización corresponde a una de las características importantes que se ha tratado de potenciar en el diseño ya que dentro de las consideraciones efectuadas, se decidió incorporar minería gráfica como una forma de potenciar los proceso de análisis por medio de herramientas visuales que guíen al analista en el proceso de descubrimiento, obteniendo vistas enriquecidas que le permiten contemplar un conjunto de características que forman parte de los textos en una sola representación gráfica.

4.4.3.1. Análisis de Frecuencias

Para determinar las características de los tokens según el tipo de datos, se diseñó una herramienta, programada en Python que recibe el nombre de Textanalizer. Esta herramienta no realiza análisis semántico o sintáctico; su objetivo es tokenizar el texto no estructurado presente en las descripciones y posteriormente analizar el comportamiento de los datos, entregando información útil para vislumbrar las características que presentan los tokens que conforman la descripción de mercancías, y al mismo tiempo determinar si se cuentan con los elementos mínimos necesarios para realizar los posteriores procesos de clasificación.

La Figura 16 muestra las siguientes tareas que se encuentran vinculadas a este análisis:

1. **Pre-procesamiento.** Durante el pre-procesamiento se obtienen los tokens que componen las descripciones y se pasan a una forma normalizada, tal como se detalló en la sección anterior.

2. **Conteo de tokens.** Se cuentan los tokens a nivel global (todo el conjunto analizado) y a nivel de registros (los tokens que forman parte de una descripción). Dentro del conteo también se clasifican los tokens en Alfabéticos, Numéricos o Alfanuméricos.

3. **Segmentación.** La segmentación por tipo de datos es utilizada en las operaciones de conteo, identificación de máximos - mínimos y análisis de frecuencia. La segmentación establecida, es la detallada en el punto anterior.

4. **Conteo de frecuencias.** Se realiza el cálculo de frecuencias tanto a nivel global como a nivel de registros. Se pretende identificar las palabras más utilizadas y al mismo tiempo se cuenta con suficientes palabras útiles que puedan ser utilizadas como atributos para permitir la individualización de los productos que pueden ser recuperados a partir de las descripciones presentes en cada transacción de importación.

5. **Identificación de vocabulario.** El vocabulario corresponde a las palabras normalizadas que no son stopwords y que no se encuentran repetidas. Con el vocabulario se puede tener una idea de la cantidad de palabras que se utilizan en las descripciones de una partida arancelaria. En el futuro, con la información recopilada, se podrá comparar vocabularios para distintas partidas y ver que tanto varían.

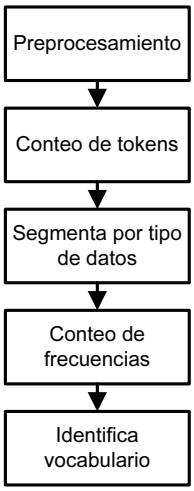

Figura 16: Tareas de TextAnalizer

Por último, dentro del conteo de frecuencia también se evalúan las frecuencias de N-gramas. Para esto se utiliza un programa que segmenta el texto en unigramas, bigramas y trigramas realizando un análisis de las frecuencias que entrega la puntuación de los N-gramas para cada una de las categorías mencionadas.

4.4.3.2.Análisis de Vinculación

El análisis de vinculaciones se lleva a cabo por medio de la creación de una red semántica, donde cada uno de los términos que aparecen en una descripción son llevados a una representación gráfica, en la que se puede apreciar la frecuencia de estos y la fortaleza de las relaciones que se establecen entre tokens.

La potencialidad de los análisis de vinculaciones fue ampliamente descrita en el estado del arte, y son llevadas a la práctica por medio de la visualización de redes semánticas y gráficos circulares, que son generados por componentes del *framework* que se encargan de crear representaciones matriciales de coocurrencias, las cuales son convertidas a una estructura de grafo dirigido para poder realizar un análisis sobre la red generada.

Para obtener la representación gráfica se tienen que realizar un conjunto de pasos, esquematizados en la Figura 17, donde se encuentran presentes las siguientes tareas:

1. **Concatenación de transacciones:** Las transacciones se concatenan en un archivo de texto donde cada registro es representado por una fila dentro del archivo. Esta tarea ya fue mencionada en el pre-procesamiento pero se referencia nuevamente ya que es un proceso compartido para ambas etapas.

2. **Representación matricial:** A partir del archivo que contiene las transacciones de importaciones junto a la descripción de productos, se procede a identificar los distintos tokens que forman parte de las descripciones, generando una matriz con las relaciones que se presentan entre estos. En la Figura 17, los

tokens están asociados a términos que se representan por la letra "T" seguido de un número correlativo, donde se define una matriz de adyacencia obtenida a partir de las relaciones que existen entre los distintos términos.

3. **Creación de la red:** A partir de la matriz que representa las coocurrencias se genera una estructura de grafo dirigido, la cual es exportada por medio de la generación de un archivo GML (estructura de archivo que permite almacenar una representación gráfica en base a texto).

4. **Inspección de la red:** La red obtenida puede ser importada desde el archivo obtenido en el paso anterior a la herramienta Gephi, que permitirá realizar la exploración del grafo generado, el cual representa la red semántica obtenida a partir de las frecuencia de los términos que forman parte de las descripciones y de las relaciones que existen entre estos.

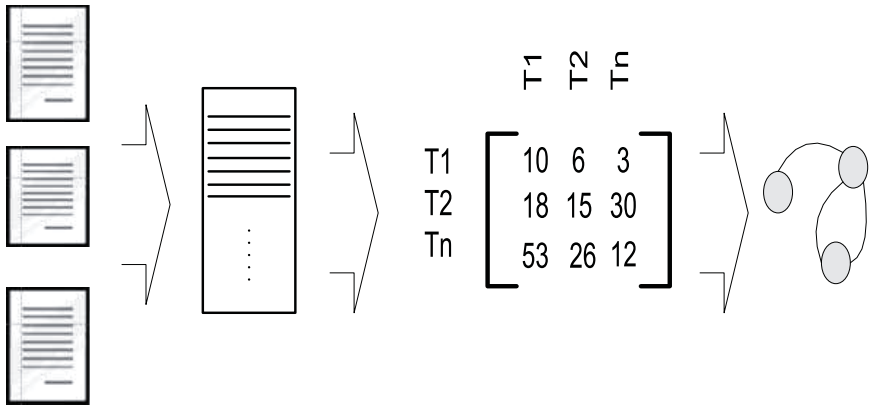

Figura 17: Transformación de datos para visualización

4.4.3.3. Selección de Características

El proceso de selección de características, considerado como parte de la extracción de palabras relevantes del texto, fue evaluado desde distintos frentes. Una de las primeras aproximaciones se basó en la utilización de las frecuencias de las palabras, mecanismo que puede ser considerado como uno de los enfoques más básicos para la extracción de características y por lo mismo, no libre de limitaciones ya que su rendimiento, según el contexto en el cual se aplica, puede llegar a ser modesto. Otra opción más avanzada corresponde a la utilización de pesos basados en *TF-IDF*, tal como se vio en el estado del arte, pero se consideró que esta técnica escapa a la naturaleza del problema ya que los textos no son complejos ni extensos y pueden ser abordados con un enfoque más simple.

También se probó con la incorporación de otros elementos al análisis de las frecuencias. Tal es el caso de los resultados obtenidos con el algoritmo *apriori* (algoritmo para la generación de reglas de asociación) y la utilización de la información provista por un análisis con *dendogramas a* partir de la aplicación de agrupamientos, mejorando de esta forma el reconocimiento de las palabras relevantes. A pesar que lo anterior presenta mejoras significativas a un enfoque basado sólo en la frecuencia, se optó por buscar un mecanismo que fuese más acorde a la esencia del

framework, en cuanto a sus características de apoyo visual, y análisis por medio de la utilización de grafos. Fue por esta razón que la solución más idónea para mantener la coherencia del desarrollo, tenía que recaer en una técnica que pudiera aprovechar la potencialidad del análisis de redes, para que por medio de la estructura previamente definida se pudieran recuperar las características relevantes. Por lo tanto, se optó por utilizar el algoritmo *PageRanking* de Google [94], el cual encajaba perfectamente en la filosofía de diseño y entregaba un método efectivo para la identificación de características relevantes, por medio de un ranking obtenido a partir de las relaciones que presentaban los términos en la red semántica.

Pero el sólo hecho de utilizar el algoritmo directamente no se considera una solución apropiada para la extracción de palabras relevantes, ya que éste se basa principalmente en la puntuación de los nodos en función de la referencia que reciben desde otros nodos. Por consiguiente, se agregó una modificación al algoritmo donde se adiciona a su fórmula un peso por nodo que corresponde a la frecuencia que tiene la palabra en el texto (las palabras no son stop words). Con la mejora incorporada, se cuenta con un mecanismo de extracción de características que se centra tanto en la importancia de la palabra en base su frecuencia de aparición en el texto, como en la puntuación que ésta pueda recibir por medio de las referencias que recibe por otras palabras dentro de la red semántica.

4.4.3.4.Clasificación y Atenuación del Ruido

Para clasificar las manifestaciones más recurrentes de ruido fue necesario desarrollar un programa que se encargara de identificar aquellas palabras desconocidas, para posteriormente determinar si se trataban de palabras que reconocía el sistema o correspondían a algún tipo de distorsión sobre los datos. Para realizar esta tarea se utilizan los diccionarios que fueron confeccionados con los atributos más sobresalientes de las descripciones, obtenidos como producto de una extracción de características de contexto, que concluye con el almacenamiento de los términos normalizados en los diccionarios. Durante el análisis de las distorsiones se trabaja a nivel de tokens alfabéticos ya normalizados, los que se comparaban con el contenido de los diccionarios, que formaban parte de la base de datos del conocimiento que forma parte del *framework*, apartando aquellos tokens que son desconocidos ya que corresponden a palabras que pueden presentan algún tipo de distorsión. Posteriormente a partir de las palabras seleccionadas como desconocidas se procede a clasificarlas según tipos de distorsión encontrada, proceso que se realizó en forma manual donde se identificaron las siguientes categorías de ruidos:

1. **Caracteres repetidos**: Corresponden a los errores donde un toquen puede estar compuesto por un carácter repetido varias veces en cualquier parte de la palabra.
2. **Símbolos y signos de puntuación:** Dentro de las descripciones analizadas existen símbolos que se utilizan para la descripción, los cuales en su mayoría recaen en unidades de medidas como las pulgadas, los porcentajes asignados a la composición de las prendas de vestir y los signos de puntuación, donde la gran mayoría a veces trata de cumplir la función de delimitador de campo, algo que en la práctica no funciona ya que no todos los agentes de Aduana tratan de delimitar campos.
3. **Palabras fusionadas:** Corresponde a palabras que se encuentran concatenadas.

4. **Palabras fragmentadas:** Es el caso inverso al anterior; una palabra es fragmentada por medio de un corte aleatorio que la transforma en dos palabras que son catalogadas como distintas.

Una vez que se tiene claro el tipo de ruido que está presente en los datos, es posible continuar con el paso siguiente, que comprende la definición de una estrategia para enfrentar el problema del ruido.

Según el tipo de ruido la complejidad varía para poderlo atenuar, por esta razón se han definido 3 enfoques para poder enfrentar las distintas categorías:

1. **Expresiones Regulares:** Las expresiones regulares permiten definir patrones que son de gran utilidad para clasificar y al mismo tiempo realizar un tratamiento sobre los datos. Esta técnica puede ser de gran utilidad para enfrentar las categorías de ruido 1 y 2.
2. **Heurísticas:** Se definió una heurística basada en que si se tienen dos palabras consecutivas, si ambas son desconocidas para el sistema la fusión de estas puede generar una palabra conocida. Esta simple heurística es utilizada con excelentes resultados en los problemas de ruido asociados al tipo 3.
3. **Modelo para corrección de lenguaje:** Para remediar los problemas del tipo 4, que resultan ser los más complejos, se generó un modelo de lenguaje para el español especializado en el contexto, que es capaz de realizar correcciones ortográficas complejas como las asociadas a esta categoría de ruido.

Las expresiones regulares y la heurística son las menos difíciles de implementar, ya que en el primer caso sólo fue necesario definir los patrones adecuados para identificar tokens con las características que se desean aislar, y en el segundo caso sólo se necesitó fusionar los tokens que cumplen con ser desconocidos, para luego comparar el resultado con los diccionarios del sistema. Pero el modelo de lenguaje presenta una complicación adicional ya que es necesario ajustar su aprendizaje a un nivel que permita reconocer la mayoría de las palabras más frecuentes utilizadas en las descripciones, con la complejidad léxica que eso puede implicar. Para solucionar esto se optó por reducir la dimensionalidad de las palabras que se encuentran en las descripciones, aplicando la normalización de los datos y la eliminación de las inflexiones.

4.4.4. Creación de Diccionarios

La descripción de productos se caracteriza por poseer un vocabulario asociado al contexto de cada producto, donde se utiliza con frecuencia una cantidad limitada de términos para realizar las descripciones. Lamentablemente la creación de un diccionario no está libre de problemas ya que la dimensionalidad de las palabras se ve afectada por las inflexiones del lenguaje, abreviaciones, sinónimos, errores ortográficos y la presencia de ruido. Por esta razón, es necesario contar con una base de conocimiento que permita apoyar a la eliminación del ruido, y al mismo tiempo sea de utilidad para disminuir la dimensionalidad de cada contexto.

Una primera aproximación para solucionar este problema fue incorporar un diccionario, lo más completo posible, en español, el cual podía ser complementado por información de contexto; pero esto demostró no ser una solución muy

eficiente ya que a medida que el diccionario poseía una mayor cantidad de palabras, incluyendo las inflexiones, éstas tendían a coincidir con algunos errores de digitación, perdiendo de este modo la utilidad del diccionario para filtrar las palabras erróneas.

Considerando lo aprendido en el párrafo anterior, se optó por definir diccionarios de contexto, los cuales serían generados con una mínima intervención humana, a partir de los datos relevantes que se puedan recuperar desde las descripciones. La generación de los diccionarios corresponde a un proceso que identifica las palabras mas frecuentes en las descripciones (sin considerar stopwords), a las cuales se les termina eliminando las inflexiones para agregarlas al diccionario. Ya que el diccionario no contempla inflexiones, se tendería a pensar que no se podrá reconocer gran parte del texto que viene en las descripciones, pero esto no es así ya que los diccionarios se ocupan en las tareas finales del pre-procesamiento, cuando los datos se encuentran normalizados y por lo tanto no presentan inflexiones.

La Figura 18 muestra el proceso por medio del cual se generan los diccionarios a partir de las características relevantes que se pueden encontrar en las descripciones.

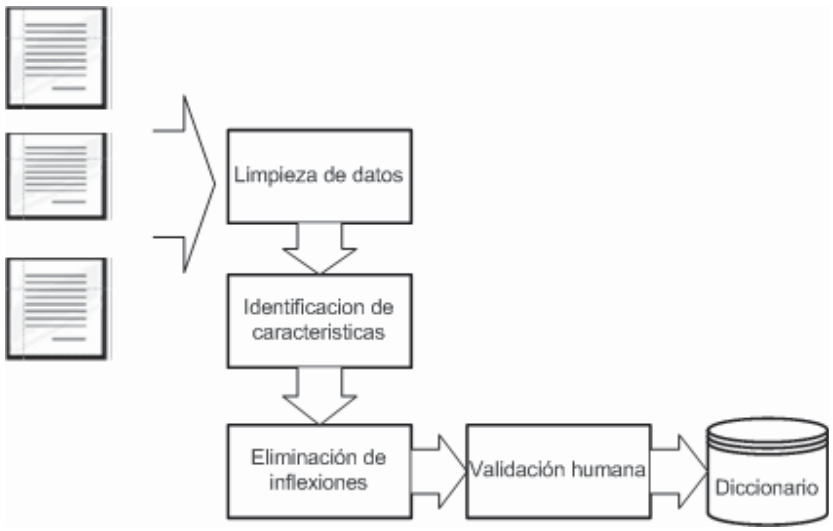

Figura 18: Creación de diccionarios

Dentro de los diccionarios disponibles se encuentra el diccionario de contexto, el cual contiene aquellos términos propios del contexto lingüístico asociado a la descripción de los productos y se caracteriza por presentar una dinamicidad mayor, característica que lo diferencia de los restantes. Por su naturaleza dinámica, el diccionario necesita de una mantención constante, orientada actualizar constantemente la base de términos en base a la aparición de nuevos elementos que presenten alguna ocurrencia significativa en las nuevas transacciones.

Los diccionarios que componen las bases de conocimiento del sistema son las siguientes:

1. **Lemario:** Corresponde a un diccionario con 360 entradas, de palabras que forman parte del vocabulario en

español.

2. **Contexto:** El diccionario de contexto contiene 174 términos asociados al área contextual en la cual se realizan las descripciones, los cuales fueron seleccionados a partir de los términos relevantes identificados.
3. **Stop Words:** Contiene un total de 428 palabras que son consideradas stop words en el idioma español.
4. **Medidas:** Corresponde a abreviaciones de las unidades de medidas más utilizadas en la descripción de productos.

4.4.5. Extracción Estructurada de Descripciones

Esta etapa se enfrenta como un proceso de extracción de información, similar a las tareas de reconocimiento de entidades. La diferencia radica en que en un proceso de identificación de entidades se recuperan aquellas palabras que reiteradamente aparecen juntas, lo que es menos complejo que identificar pares atributo - valor ya que un atributo puede tener una gran cantidad de valores que están asociados a él, lo que eleva la complejidad del proceso de extracción.

Después de contar con el conocimiento previo sobre las características de los datos, se pueden identificar cuatro situaciones a las cuales el *framework* tiene que enfrentar:

1. La primera es la más simple y corresponde a la extracción de atributos que son de la forma (valor - unidad de medida); en estos casos es fácil recuperar un par ya que al identificar un token que corresponde a una unidad de medida, cercano a ésto se tiene que encontrar el que corresponda al valor de dicha unidad. Un ejemplo de esto son los tokens que representan la composición de las telas.

2. El segundo caso corresponde a los atributos binarios, cuya presencia indica por si misma la existencia de una característica. Un ejemplo de esto es el atributo “color” ya que si esta palabra figura, el atributo es verdadero, lo que clarifica que la ropa es de color.

3. Un tercer caso corresponde a la identificación de atributos recurrentes pero que pertenecen a un conjunto muy limitado. Un ejemplo son las marcas, las cuales pueden ser identificadas por un diccionario que se obtenga a partir de la recuperación de marcas por medio de la frecuencia, y de la relación de éstas con palabras claves que ayudan a su identificación. Todo lo anterior se puede hacer por medio de algún procedimiento no supervisado con una mínima intervención humana.

4. El cuarto es el tipo más complejo, un atributo está compuesto por un conjunto de palabras al igual que los valores que pueda tomar. Para enfrentar esto es necesario identificar las palabras que ocurren juntas y conocer qué parte de la frase formada por ellas es la que varía en un conjunto de de tokens acotados, los cuales corresponden al valor que puede tomar el atributo.

Para enfrentar estas situaciones se ha optado por abordarlas de dos maneras distintas. Para identificar los atributos que no son compuestos, se ha decidido utilizar un etiquetador diseñado para identificar elementos del texto que

ayuden a extraer un par atributo valor, y para el caso más complejo que corresponde a la situación 4 se ha optado por utilizar un clasificado bayesiano.

En ambos enfoques se utiliza la información de coocurrencias obtenida a partir de la red semántica que se crea en las etapas preliminares del análisis.

A continuación se describe la forma de trabajo de estos dos enfoques.

4.4.5.1.Extracción Utilizando un Etiquetador.

Esta alternativa es útil para reconocer elementos muy recurrentes y de fácil clasificación, viniendo a solucionar casi la totalidad del problema de recuperación de atributos en la partida T-Shirt ya que sus atributos relevantes presentan este tipo de representación. En este caso considerar la utilización de un etiquetador, como apoyo para realizar una extracción de atributos en base a reglas, puede ser una herramienta útil para trabajar sobre conjuntos de datos acotados y conocidos. En el caso de la partida arancelaria motivo de estudio, esto es factible de aplicarlo para los atributos que definen marcas, composición de la tela, color y logotipos, cubriendo gran parte de la clasificación de atributos que pueden ser considerados relevantes para identificar sub-valoraciones.

Para crear este componente se procedió a enfrentar el problema como si se tratara de un etiquetador POS, pero en vez de identificar las partes de una oración correspondiente a un análisis morfológico, el etiquetador se preocupa de identificar elementos conocidos dentro de una descripción.

Como existen variados enfoques que van desde la utilización de diccionarios, identificación de patrones en base a expresiones regulares y técnicas de aprendizaje automatizado para diseñar un algoritmo POS, se optó por utilizar lo aprendido en el diseño de estos algoritmos, para crear un etiquetador de descripciones. De esta forma, una vez que se tiene el texto etiquetado con los elementos de las descripciones que reconoce el etiquetador, se puedan aplicar reglas de recuperación para obtener los pares atributos valor, utilizando como ayuda la información provista por las etiquetas.

En el diseño del etiquetador se utilizaron los recursos que entrega la API NLTK para crear etiquetadores POS en Python, donde además se utiliza la característica que permite complementar distintos enfoques, con el fin de potenciarlo por medio del trabajo conjunto de distintas técnicas.

Como resultado se obtuvo un etiquetador que requiere de un corpus etiquetado para su entrenamiento, el cual fue creado manualmente en base a la información provista por la etapa de análisis de datos, sobre las descripciones de los productos de Aduana. Con el corpus etiquetado creado, se entrenó el etiquetador para obtener un modelo de clasificación, el cual fue complementado por un etiquetador por defecto, que clasifica todo lo que no reconoce con una etiqueta de desconocido; más un etiquetador en base a diccionarios, donde están contenidas las marcas y las unidades de medidas que se utilizan en las descripciones.

Con esos tres componentes se pudo diseñar un etiquetador compuesto, que identifica cada elemento significativo de las descripciones, permitiendo posteriormente aplicar reglas de recuperación en base a la identificación de pares de etiquetas conocidas o por el análisis de coocurrencia de las palabras etiquetadas.

Es necesario agregar como dato de interés, que un algoritmo POS fue descartado del *framework* ya que un análisis morfológico no tiene sentido en el contexto de las descripciones de importaciones. Esto debido a que se descubrió que los elementos que las componen son todos "nombres", los cuales están seguidos de elementos del mismo tipo o de tokens numéricos. Identificar verbos u otros componentes de la oración, es una tarea que no aporta mayor valor en este caso, debido a los limitados elementos que se pueden distinguir en la composición de las oraciones descriptivas, la creación de un etiquetador para identificar estructuras que se observan en las descripciones parece ser la solución más idónea para enfrentar la recuperación de pares atributo-valor, donde el nivel de complejidad permite una sencilla clasificación.

4.4.5.2.Extracción Utilizando un Clasificador Bayesiano.

Para los casos donde el etiquetador y la extracción en base a reglas no son suficientes, se utiliza el enfoque propuesto en los trabajos de Rayid Ghani y Katharina Probst, donde se plantea la utilización de datos etiquetados y sin etiquetar en un entorno de aprendizaje semi-supervisado para extraer pares atributo-valor desde texto no estructurado, proceso que fue descrito con anterioridad en el estado del arte.

Al igual que en los trabajos antes mencionados, se ha optado por considerar la tarea de extracción de atributos y valores como una problema de clasificación, en el cual se pretende clasificar las palabras/frases como atributo, valor o ninguno de los anteriores.

Si bien es cierto, el enfoque coincide con los trabajos estudiados, al realizar la recuperación de atributos utilizando un clasificador bayesiano, por otro lado difiere en las técnicas restantes, principalmente porque se consideró irrelevante la utilización de algoritmos POS por las razones expuestas en el apartado anterior. Además se optó por simplificar la etapa de aprendizaje, acotando la solución por medio de la utilización de un algoritmo supervisado, entrenado para reconocer un conjunto limitado de pares atributo-valor, los cuales fueron seleccionados con anterioridad por medio de los resultados obtenidos en el proceso de extracción de características. De este modo el problema queda acotado, dejando para trabajos futuros la incorporación de un entorno de aprendizaje semi-supervisado por medio de co-entrenamiento.

El *framework* cuenta con un clasificar bayesiano que se encarga de clasificar cada token como atributo, valor o ninguno de éstos. Los atributos y valores recuperados se obtienen en forma separada sin identificar en esta etapa relaciones en los pares atributo-valor, ya que la tarea de clasificación sólo los encasilla, no los relaciona.

Una vez que se cuenta con la salida del proceso de clasificación (atributos y valores sin asociar), se procede a agrupar los pares atributo-valor para lo cual se consideró la utilización de algunas heurísticas en conjunto con la

puntuación de coocurrencia que presentaban los tokens. De esta forma se asocian los que presentan una mayor puntuación, relacionada a su aparición conjunta.

Tal como se ha podido apreciar en este apartado y en los anteriores, la utilización de las coocurrencias es un elemento fundamental en el diseño del *framework*, los que toman un papel relevante desde las etapas iniciales del análisis hasta la etapa de clasificación.

4.4.6. Minería de Datos

Una vez que se cuenta con el subconjunto de atributos seleccionados para los registros de importaciones de la partida T-Shirts (ver Tabla 6), más una representación estructurada de los productos, obtenida como parte de la aplicación de técnicas de minería de texto, es posible avanzar al diseño de la última etapa del *framework*, la cual comprende las operaciones de minería de datos.

En la etapa de minería de datos se recibe las descripciones de los productos en forma estructurada y se utiliza esta información para crear agrupamientos, con el fin de obtener un identificador para todos aquellos productos con atributos semejantes.

Para la detección de transacciones fraudulentas se recurre a la utilización de un análisis de anomalías. Desde la perspectiva del análisis de datos, los casos que puede corresponder a un tipo de fraude, corresponden a aquellos que presentan una desviación "importante" sobre los que son considerados normales. Si se asume que la distribución de los precios unitarios de los productos se puede presentar como una distribución normal, la identificación del precio unitario que mejor represente al valor del producto, puede ser obtenido en base a las agrupaciones de mayor densidad que contienen los valores más comunes, siendo considerados *outliers* aquellos valores que se aparten de estas.

Si se utilizan técnicas de agrupamiento no supervisados, ya que se desconoce el posible valor de precio unitario para cada producto, se cuenta con un análisis por medio de una técnica descriptiva para identificar las distribuciones presentes en los datos y en base a estas agrupaciones detectar los objetos que poseen un comportamiento anormal. Para lo anterior es posible utilizar un algoritmo de agrupamiento como K-mean o K-median, pero producto de las limitaciones de Clementine 12, sólo se puede utilizar K-mean.

Una vez que se tiene definidas las fronteras entre los agrupamientos que se puedan obtener, el siguiente paso es determinar el *ranking de outliers*. Para esto último se utiliza el algoritmo LOF, el cual se aplica sobre los datos pertenecientes a cada agrupamiento. LOF ha sido seleccionada por trabajar en base a densidades, lo que está relacionado al razonamiento antes planteado para la detección de anomalías.

Además, estas agrupaciones de productos también pueden ser utilizadas en conjunto con la información de los campos relevantes de las transacciones de importaciones, para obtener patrones de comportamiento en base al estudio de las diversas variables seleccionadas.

Dentro de los tipos de salida que genera el flujo de trabajo diseñado en de Clementine, se pueden destacar las siguientes:

- Frecuencia de transacciones por agentes
- Frecuencia de transacciones por país origen
- Transacciones por puerto de origen
- Tabulación de los resultados de los agrupamientos
- Representación gráfica de los clusters

Según lo aprendido en el desarrollo de la tesis, la flexibilidad y potencia del entorno Clementine ha sido de gran ayuda para realizar los prototipos de minería de texto; al mismo tiempo ha demostrado ser una herramienta amistosa que facilita la comprensión de las tareas realizadas, y las transformaciones que se aplican sobre los datos por medio de la representación gráficas de éstas, donde en el flujo de trabajo se puede apreciar claramente la interacción entre los componentes. Estas características también han sido de gran ayuda para presentar los resultados ya que facilita la comprensión del proceso en su totalidad, por medio de la inspección paso a paso de las etapas que lo componen. Por esta razón no resulta extraño ver la misma forma de trabajo replicada en herramientas open source como RapidMiner y Orange, entre otros, donde además de la potencialidad provista por el soporte de distintas técnicas de minería de datos, ahora también se preocupan por obtener una interfaz amistosa en base a flujos de trabajo.

4.5. DESCRIPCIÓN DETALLADA DEL FLUJO DE TRABAJO

Una vez descrita las macro etapas y los elementos que las componen, es posible realizar una representación más detallada de los componentes que conforman el *framework*. A continuación se realiza una descripción del *workflow* implícito en el *framework,* para ejemplificar la interacción que existen entre los componentes que lo conforman y las transformaciones que sufren los datos a medida que son procesados por las distintas etapas. El objetivo que se persigue es sólo realizar una descripción a nivel de flujo de trabajo, en base al diagrama presentado en la Figura 19, ya que en la sección siguiente se realiza una presentación detallada sobre los resultados sobresalientes obtenidos en las etapas más relevantes.

E0. Por medio de consultas SQL se recupera la información de las transacciones de importaciones (DIN) desde una base de datos histórica. Se genera como salida un registro compuesto por los campos que se muestran en la Tabla 6, donde el contenido del campo "descripción", corresponde a una fuente no estructurada, que contiene los datos de entrada que serán analizados por las tareas de minería de texto.

E1. Corresponde al pre-procesamiento de los datos, recibe como entrada el texto no estructurado de las descripciones, el cual es transformado en tokens. Sobre los tokens obtenidos se eliminan las stopswords y posteriormente se procede a normalizar los elementos restantes, transformando los strings a minúsculas, eliminando los acentos y por último, tratando las inflexiones presentes en las palabras. Para realizar una tokenización adecuada, esta tarea se retroalimenta de la información que entrega la etapa de análisis previo.

E2. La etapa de análisis previo se encarga de realizar el estudio de los datos no estructurados para descubrir patrones en éstos. También es utilizada para evaluar los avances que se obtiene al aplicar las técnicas de pre-procesamiento, normalización y de atenuación del ruido presente en los datos. Las tareas que la componen son las siguientes:

E2.1. Se realiza el análisis de las frecuencias por tipos de datos, palabras y n-gramas; para esto, se desarrolló un conjunto de programas que se encargan de transformar el texto en tokens, para luego realizar la evaluación de frecuencias partiendo por unigramas hasta llegar a trigramas. Producto de este análisis se obtiene el vocabulario que se utiliza en las descripciones junto a las palabras más relevantes, estas últimas son de gran ayuda para la selección de las palabras claves.

E2.2. Corresponde a la inspección de las posibles relaciones que existen entre los tokens para descubrir patrones que permitan determinar las asociaciones atributo valor presente en las descripciones; en esta tarea se agrupan los tokens a nivel de n-gramas y se identifican las co-ocurrencias más significativas. Para facilitar el análisis de vinculaciones y permitir una mejor interacción con los analistas, el *framework* entrega representaciones graficas por medio de un grafo dirigido, donde los nodos corresponden a los tokens y la relación entre éstos se representa por medio de los arcos que los vinculan.

E2.3. Corresponde a una tarea de inspección que utilizando el conocimiento de las palabras que presenta mayor frecuencia permite identificar palabras desconocidas y por medio de la inspección de éstas determinar si pueden ser clasificadas en alguna categoría de ruido. La salida de la tarea corresponde a la identificación de las clases de ruido más frecuentes, donde el tratamiento de las distintas manifestaciones de ruido, puede ser significativo para mejorar la calidad de los datos.

E2.4. La selección de palabras claves se realiza por medio de un proceso de selección de características. Este proceso se basa en el conocimiento adquirido por las etapas anteriores, principalmente en los análisis de frecuencias y coocurrencias. A partir de estos datos, se genera una estructura de grafo que permite complementar las frecuencias junto con los resultados del algoritmo *PageRanking,* para realizar una selección de características..

E3. Representa los recursos lingüísticos de carácter persistente, que utilizan tanto las tareas de pre-procesamiento como las que son parte de la supresión de ruido. Dichos recursos corresponden a distintos diccionarios que han sido agregados paulatinamente a medida que se afinan los resultados que se obtienen desde el bloque de minería de texto. La creación de los diccionarios implica la utilización de procesos que permiten incorporar nueva información de contexto a medida que aparecen nuevos datos relevantes. Toda la

información presente en los diccionarios corresponde a datos normalizados para disminuir la dimensionalidad de éstos.

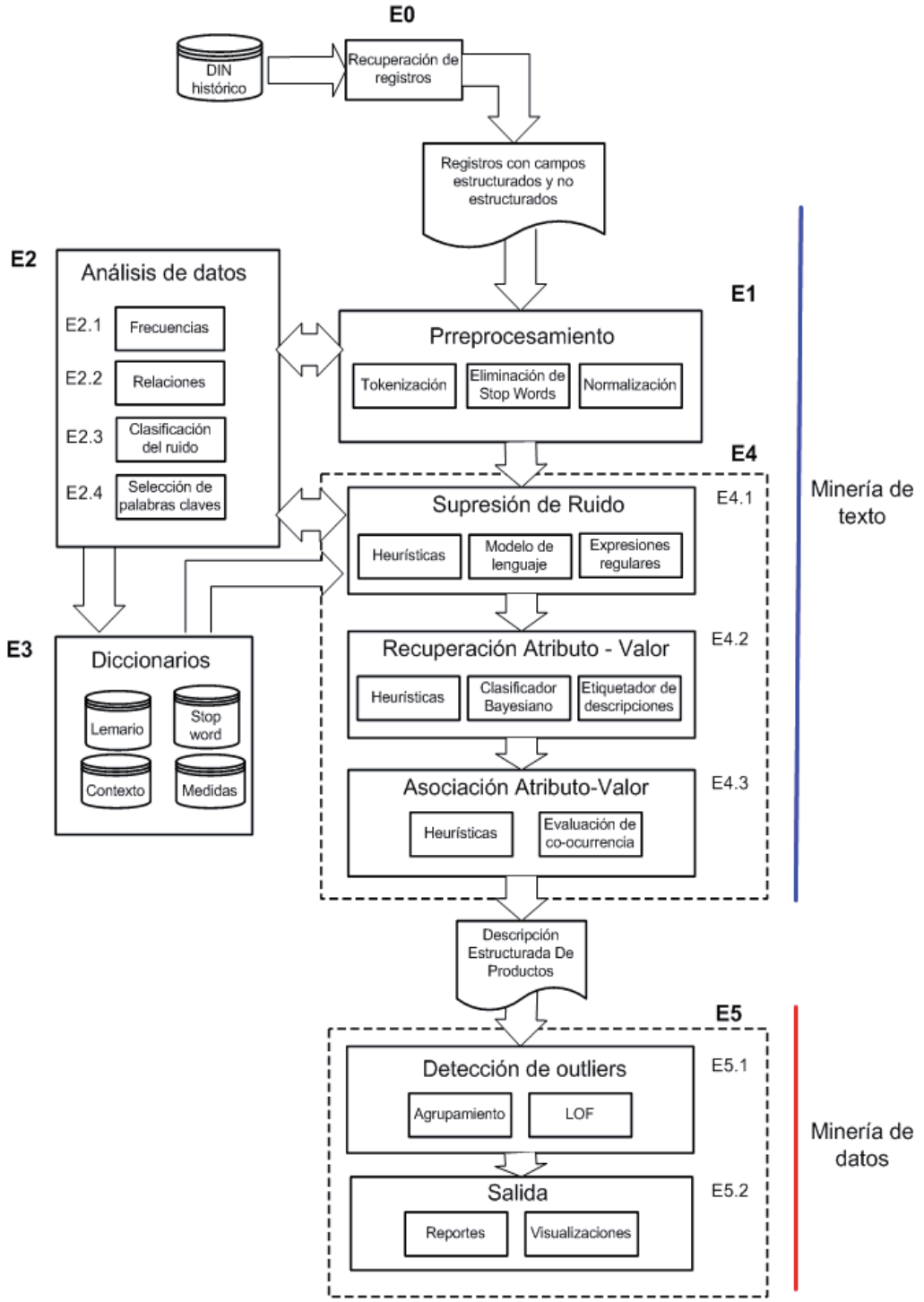

Figura 19: Workflow del Framework

E4. Incluye las tareas que forman parte de la extracción estructurada de descripciones: atenuación del ruido y de extracción de información, éstas últimas apoyadas por los modelos para el tratamiento de lenguaje. Además, las tareas se caracterizan por utilizar variados recursos lingüísticos para el cumplimiento de sus objetivos. A continuación se describen brevemente cada una de ellas.

E4.1. Recibe como datos de entrada los resultados en forma de tokens normalizados, obtenido en la tarea anterior, y se procede a disminuir el ruido que presentan estos. Dependiendo del tipo de ruido que se identifique se aplican un tratamiento por medio de heurísticas, expresiones regulares o en los casos más complejos se procede a utilizar un modelo de lenguaje para realizar correcciones ortográficas y tratamiento de tokens fusionados.

E4.2. Una vez que se ha logrado atenuar parte del ruido producto de las tareas ejecutadas en la etapa anterior, es posible continuar con la clasificación de productos, la cual se dividen en dos frentes:

1.- Los atributos de los productos pueden ser obtenidos por un etiquetador que se encarga de identificar aquellos componentes de las descripciones que son considerados de utilidad para posteriormente extraer pares atributo-valor. Como salida el etiquetador entrega un texto enriquecido con tags para utilizarlos en la etapa siguiente como ayuda para la extracción.

2.- El otro camino consiste en utilizar un clasificador bayesiano encargado de diferenciar los tokens que corresponde a la clase atributo, valor u otros; por cada descripción se obtiene un conjunto de atributos y valores por separado. En esta etapa aún no existe una relación entre los tokens previamente clasificados, sólo se sabe que por medio de la discriminación realizada por el clasificador, los tokens recuperados se encuentran dentro de las tres categorías (atributo, valor, otros), de las cuales se descarta la última categoría en el proceso siguiente. Es importante destacar que la clasificación se basa en una plantilla de atributos que han sido seleccionados con anterioridad, los cuales corresponden a un subconjunto conocido, a partir de la selección de características realizadas en la etapa **E1**. Este método es utilizado para clasificar atributos más complejos donde se hace necesario utilizar coocurrencias para identificar conjuntos de palabras que pueden formar parte de un atributo o valor.

E4.3. Por último, sobre los conjunto de atributos y valores, se procede a realizar asociaciones entre los tokens que pueden ser parte de un atributo o valor, para posteriormente sobre dicho tokens realizar el pareo atributo-valor, guiado por las puntuaciones de coocurrencia disponibles.

E5. Corresponde al bloque de minería de datos, el cual recibe como entrada un conjunto de datos, compuestos por los campos estructurados definidos en la Tabla 6, más la información que ha estructurado el proceso de minería de texto por medio de la recuperación de pares atributo-valor a partir de los campos descriptivos.

E5.1. En el proceso de minería de datos se utiliza toda la información estructurada para realizar distintos análisis sobre los datos, donde la técnica más utilizada es el agrupamiento. Por este medio se pueden segmentan aquellas transacciones que presenten comportamiento extremos a la norma, la cual es obtenida a partir de los datos con características semejantes. Después de utilizar el agrupamiento para identificar datos homogéneos, se realiza una detección de *outliers* por medio de la aplicación del algoritmo LOF.

E5.2. Por último, los agrupamientos pueden ser visualizados por medio de representaciones gráficas para facilitar la navegación y comprensión de los resultados.

4.6. CONCLUSIONES

Durante el diseño se han expuesto las técnicas seleccionadas y las decisiones de diseño que justifican la utilización de los elementos que componen el *framework*. En el capitulo siguiente se expondrán aquellas salidas relevantes, que se obtienen con la utilización de esta herramienta sobre los datos de prueba seleccionados, donde además se exponen algunos descubrimientos realizados.

5. RESULTADOS OBTENIDOS

En este capítulo se presentan los tipos de salidas más importantes que genera el *framework*, los resultados obtenidos producto del análisis sobre los datos y se describen los descubrimientos que se obtuvieron al realizar la identificación de transacciones anómalas.

5.1. ANÁLISIS DE DATOS

Durante la etapa de análisis se realizan un conjunto de evaluaciones que tienen por objetivo conocer las características de los datos, disminuir su dimensionalidad, identificar y atenuar los niveles de ruido y seleccionar aquellas características relevantes para ser utilizadas en la clasificación y extracción de pares atributos valor. Para cumplir los objetivos mencionados, es de fundamental importancia realizar un análisis de frecuencia y de vinculaciones, cuyos resultados son expuestos a continuación.

5.1.1. Análisis de frecuencia

Los primeros resultados que se obtienen tienen relación con un análisis de frecuencia por tipo de tokens, los resultados obtenidos se detallan en las Tablas 7 y 8. Los datos que muestra la Tabla 7 indican que al analizar las transacciones del año 2010 para la partida arancelaria de T-Shirts, en los campos descriptivos de todos los productos se obtuvo un total de 129.612 tokens, de los cuales 20.404 correspondían a *stopword* reconocidas; si a los tokens totales se les resta estas palabras, da un total de 109.208 palabras. Como estos tokens pueden estar repetidos, sólo se consideran los no repetidos lo que disminuye la cantidad anterior a 15.478, entre los cuales se encuentran tokens del tipo alfabético, numérico y alfanumérico (Tabla 8). Si sólo se consideran los tokens alfabéticos se tiene el tamaño del vocabulario utilizado en la descripción de T-Shirts, lo que arroja un total de 1.847 palabras distintas. Es importante aclarar en este punto, que el tamaño del vocabulario detectado contempla inflexiones por lo tanto éste puede ser mucho menor si éstas se eliminan.

ANALISIS DE MUESTRAS GLOBALES	
Tokes totales	: 129612
Stop Word detectadas	: 20404
Tokens sin stopwords	: 109208
Tokens SIN repetir	: 15478
Vocabulario	: 1847

Tabla 7: Resultados generales para T-Shirts.

Cantidad de tokens por tipo de dato SIN incluir repetidos			
Alpha :1847	Num:94	Alphanum:13537	Total:15478

Tabla 8: Segmentación de tokens no repetidos por tipo.

Por otro lado, si se realiza un análisis de frecuencia de los tokens que aparecen en cada tipo de dato, se obtienen los resultados presentes en las Tablas 9, 10 y 11, donde las cantidades de tokens se presentan en orden descendente. Del análisis de frecuencia de los datos se puede concluir lo siguiente:

1. Las primeras mayorías de los tokens alfabéticos contienen palabras que corresponden a atributos, aunque también queda en evidencia que es necesario preprocesar las palabras que aparecen en plural, lo que corresponde a un tipo de inflexión morfológica.

2. Los valores presentes en los tokens numéricos, en su mayoría están asociados a la composición de las prendas.

3. La primera mayoría de los tokens alfanuméricos, que también comprende los signos de puntuación, recae en el carácter "**;**", lo cual tiene algo de sentido ya que el SNA pretendía que los agentes de Aduana utilizaran este carácter como delimitador de campos. Lo último no tuvo el éxito esperado, al no existir acuerdo en aplicar ésta práctica; por dicha razón los caracteres de puntuación se consideran como parte del ruido ya que no entregan valor al análisis. Caso similar es el carácter de las comillas (**"**) que sólo distorsiona los datos al generar palabras distintas por el hecho de fusionarse a éstas. Por último, se puede apreciar que un problema no menor a remediar, estará asociado a la fusión de tokens ya que esto es algo recurrente según se observa en la Tabla 11.

Número	Frecuencia	Porcentaje	Token
1	5081	15	tejido
2	2466	7	punto
3	1555	4	pleras
4	1443	4	polera
5	1252	3	tallas
6	1230	3	algodon
7	1193	3	tejidos
8	756	2	diferentes
9	694	2	manga
10	611	1	cuello

Tabla 9: Top 10 para tokens alfabéticos.

Número	Frecuencia	Porcentaje	Token
1	319	48	100
2	37	5	3
3	31	4	5
4	28	4	95
5	24	3	1
6	24	3	2
7	10	1	180
8	9	1	170
9	8	1	10
10	8	1	97

Tabla 10: Top 10 para tokens numéricos.

Número	Frecuencia	Porcentaje	Token
1	9492	12	;
2	3374	4	100%
3	2591	3	punto”
4	2033	2	100%algodon;
5	2029	2	Algodon;
6	1617	2	Mujer;
7	1387	1	Hombre;
8	1317	1	“sin-codigo
9	1279	1	punto,
10	1229	1	camisetas;

Tabla 11: Top 10 para tokens alfanuméricos.

En la Tabla 12 se puede ver que el porcentaje de las stopwords concuerda con las estimaciones que existen sobre la proporciones de stopwords en los textos (cercano al 20%), estas se encuentran incluidas dentro del porcentaje de tokens alfabéticos. Por otro lado, llama la atención el alto porcentaje de tokens alfanuméricos, pero según el análisis de frecuencia la explicación consiste en que en las descripciones se tienden a fusionar las palabras con los signos de puntuación, notaciones de unidades de medidas y en muchos casos los nombres de las unidades de medidas con los datos numéricos que representan las cantidades.

Items	T-Shirts
Stopwords	16 %
Tokens alfabéticos	12 %
Tokens numéricos	1%
Tokens alfanuméricos	87 %

Tabla 12: Resumen de resultados

La Tabla 13 contiene los resultados de los tokens identificados en las descripciones que no se encuentran repetidos. La distinción anterior, se realiza para determinar el tamaño del vocabulario utilizado, el cual se puede segmentar en palabras, números y tokens alfanuméricos, en los cuales también se incluyen las palabras fusionadas; con unidades de medidas o signos de puntuación. En los resultados expuestos en la Tabla 13, llama la atención la enorme cantidad de tokens alfanuméricos y sin repetir; esto está directamente vinculado a una alta dimensionalidad de los datos, la cual se relaciona a la presencia de una fuerte presencia de ruido en los estos.

Vocabulario	T-Shirts
Alfabético	1.847
Numéricos	94
Alfanuméricos	13.537
Tokens sin repetir	15.478

Tabla 13: Resultado del vocabulario identificado

5.1.2. Análisis de vinculaciones

Una de las potencialidades del *framework* se basa en la utilización de representaciones gráficas para realizar una inspección sobre los datos, con el fin de presentarlos en una sola vista enriquecida, en la cual se incluyen un conjunto

de valores útiles para el análisis; tales como la frecuencia, las coocurrencias y la fortaleza de las relaciones que se establecen entre los nodos que representan palabras.

La Figura 20 muestra la riqueza en contenidos que provee una visualización sobre la red semántica, donde el tamaño de los términos representa la distribución de las frecuencias sobre los datos, y la fuerza (peso) de las relaciones es representada por el nivel de intensidad de los arcos que unen los nodos.

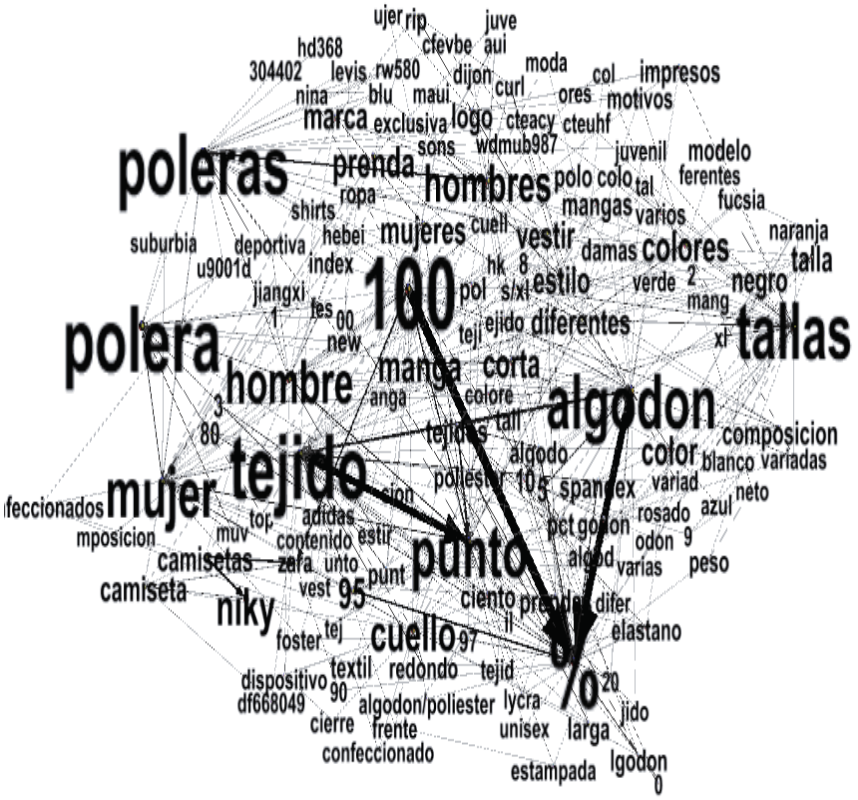

Figura 20: Red semántica para descripciones de T-Shirts

De la red semántica generada para las descripciones de la partida T-Shirt se puede inferir lo siguiente:

1. En los términos que presentan mayor frecuencia existen inflexiones que tienen que ser eliminadas. Tal es el caso de la palabra polera y su forma plural.

2. Por tratarse de un grafo dirigido se puede apreciar el sentido de las asociaciones lo que facilita la identificación de los atributos y valores más recurrentes.

3. Dentro de los atributos que más aparecen se encuentran los que están asociados a la composición de las prendas.

4. A simple vista se pueden apreciar alguna de las marcas que presentan mayor frecuencia.

5. Debido a que las palabras aún no se encuentran normalizadas, la gráfica permite identificar algunos tipos frecuentes de distorsiones y palabras distintas que tienen el mismo significado.

Es conveniente recalcar que este tipo de representación tiene por objetivo servir de apoyo para asistir al proceso de descubrimiento por medio de una inspección interactiva. Para la tarea de seleccionar un conjunto de características relevantes a partir de los datos, la inspección se realiza de manera automatizada en base a la selección de los elementos más significativos. Dichos elementos se obtienen después de limpiar los datos, realizando un análisis de frecuencia, para luego aplicar el algoritmo PageRanking.

El análisis de vinculaciones también es utilizado para validar los resultados obtenidos en etapas avanzadas del procesamiento de texto. Por ejemplo, después de aplicar las tareas de normalización es posible realizar una visualización de los resultados para ver como han variado las relaciones y la dimensionalidad de los datos. Un ejemplo de lo antes expuesto corresponde a lo presentado en la Figura 21, donde un layout distinto permite una mejor visualización de los términos relevantes. En la vista obtenida, los datos con mayor frecuencia y cantidad de relaciones se presentan en un primer plano lo que facilita notablemente su identificación. Se puede ver también, que no existen errores en las palabras, los datos han sido normalizados y producto de la utilización de los sinónimos, han surgido nuevos términos que se encuentran dentro de los de mayor frecuencia, como es el caso de los géneros, que en muchas descripciones figuraban en inglés.

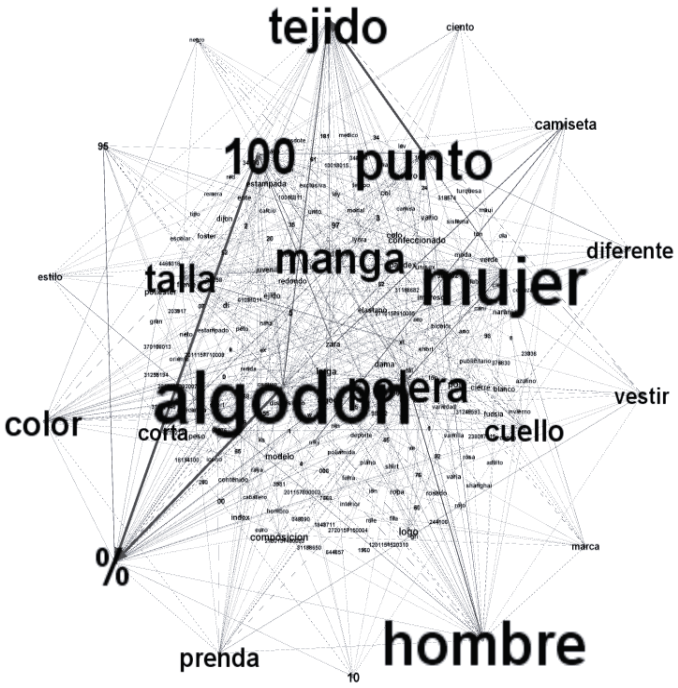

Figura 21: Red semántica con datos normalizados

Además del cambio de layout que presenta la figura anterior, existen representaciones muy útiles, las cuales por medio de un cambio de la topología, facilitan notablemente el descubrimiento de las relaciones entre los datos. Un

ejemplo de esto, es el cambio a una representación circular, la que en muchas ocasiones permite tener una mejor visualización de las relaciones que se establecen, tal como lo expone la Figura 22.

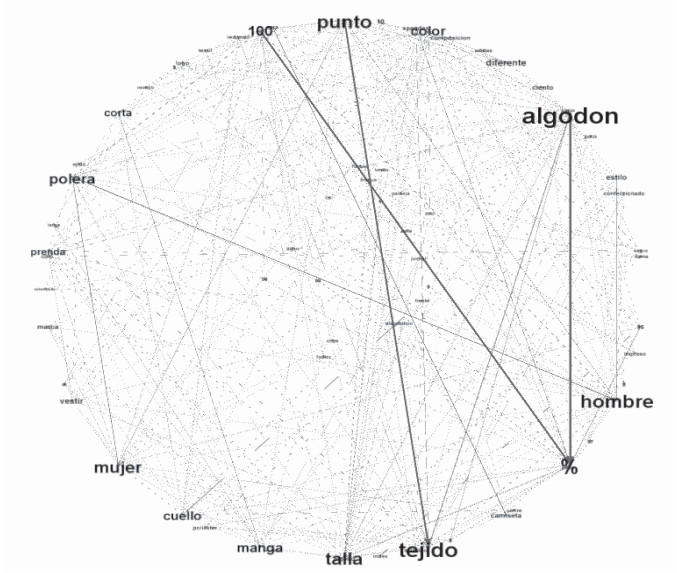

Figura 22: Gráfico circular para términos de T-Shirts

Otra tarea que es facilitada por medio de la inspección gráfica, corresponde a la identificación de algunos patrones de interés. Por ejemplo, es posible descubrir los modelos asociados a una marca y la forma utilizada para su codificación, tal como se aprecia en la Figura 23, donde aparece una marca Niky (con alcance de nombre para la marca Nike) que cuenta con asociaciones a múltiples campos numéricos, los que corresponden a la codificación que realiza un agente en particular, para codificar los distintos modelos dentro de una marca.

Figura 23: Modelos de una marca

Todas las representaciones que se han expuesto surgen gracias a las características incorporadas al *framework* para utilizar la potencialidades que entrega la *minería gráfica*, lo que permiten explorar y analizar espacialmente las frecuencias de los datos, sus relaciones y las transformaciones que ocurren sobre estos a medida que se avanza por

las etapas de procesamiento del texto. Estas características de visualización avanzada corresponden a uno de las funcionalidades sobresalientes que presenta este trabajo, la cual persigue entregar un ambiente de trabajo potente y al mismo tempo amigable e intuitivo para realizar los análisis en las etapas de minería de texto.

5.2. SELECCIÓN DE CARACTERÍSTICAS

Después del pre-procesamiento de datos, el *framework* permite realizar la selección de características en base a la utilización del algoritmo PageRanking modificado, tal como se detalló en el capitulo de diseño. Las salidas del trabajo de puntuación obtenidas pueden ser consultadas por el analista en base a la representación de las puntuaciones por medio de tablas, o utilizando una representación gráfica.

En la Figura 24 se muestra la salida de un proceso de ranking, donde aparecen las palabras seguidas de la puntuación obtenida; además se puede observar el nivel de relaciones que cada palabra presenta con las restantes. En la gráfica los resultados de puntuación son números positivos que han sido amplificados y redondeados a dos decimales para mejorar la presentación gráfica.

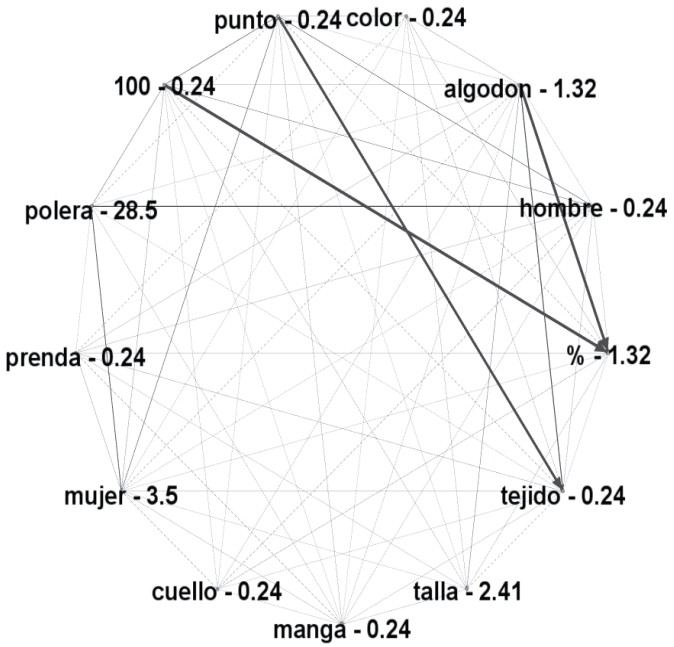

Figura 24: Puntuación de palabras

Otro aspecto importante del gráfico de la Figura 24 corresponde a la incorporación de campos numéricos y signos en la estructura. Por lo general, los valores numéricos se descartan de los análisis de texto, al igual que los signos, pero en este caso se ha evitado eliminar los signos que son relevantes en las descripciones, como el porcentaje, ya que permiten identificar los tokens que están asociados a la composición de las telas. Lo mismo ocurre para los valores numéricos ya que presentan ocurrencias puntuales, claramente identificables, por lo que se pueden considerar como parte de los elementos significativos para analizar.

Después de realizar la extracción de las características, se pudo comprobar que existen datos suficientes para clasificar a los productos que se encuentren en la partida arancelaria para T-Shirt.

A continuación fue necesario definir aquellos atributos significativos que además cumplen con la característica de aportan valor para identificar productos que presenten subvaloraciones. Para esto se utilizó la información recopilada a partir de los análisis anteriores y se definieron 5 categorías de atributos que son expuestos en la Tabla 14.

N°	Atributo	Descripción
1	Marca	Nombre del fabricante
2	Tipo	Características que diferencian a modelos de poleras
3	Composición	Corresponde a la composición de la tela
4	Color	Si la polera es de color
5	Estampado	Si la polera posee estampado

Tabla 14: Atributos de interés

Los atributos seleccionados corresponden a aquellas características de los productos que son considerados importantes para ser utilizadas en un análisis de subvaloración de productos, ya que pueden influir directamente en el precio de las prendas. En este punto se tiene que mencionar que la partida T-Shirt es una de las consideradas complejas ya que es muy difícil encontrara atributos relevantes para ser utilizados en la detección de subvaloraciones. Distinto es el caso de partidas arancelarias donde los atributos son claramente identificables, los atributos importantes aparecen casi en la mayoría de las descripciones y tienen directa influencia sobre el valor del producto. Por ejemplo, en una partida asociada a electrodomésticos como los televisores LCD, atributos como el tamaño de la pantalla, la resolución, marca y modelos aparecen frecuentemente en las descripciones, en parte porque este tipo de productos se describe con mayor prolijidad y detalle, lo que contrasta notablemente con el caso estudiado, el cual surgió como una propuesta de Aduana para ver los resultados que se podían obtener en un caso tan extremo como el indicado.

5.3. CLASIFICACIÓN DEL RUIDO

La clasificación del ruido, es un análisis que sólo se realiza una vez y tiene por objetivo identificar las distorsiones más frecuentes sobre los datos de las descripciones, para luego clasificarlas en categorías, y así definir los mecanismos necesarios para aplicar atenuadores de ruido según sea el caso.

Se han incorporado estos resultados en el presente capítulo, porque representan a una de las salidas que se obtienen producto de la evaluación de los datos, por medio de las herramientas de procesamiento de texto que entrega el *framework*. Para realizar este análisis se procedió a identificar los agentes que presentaban el mayor número de transacciones y a partir de este subconjunto, se obtuvo una muestra representativa para poder evaluar los problemas que presentan las descripciones, en cuanto a la presencia de ruido. Los resultados obtenidos son expuestos en la Figura 25; donde se puede apreciar que aproximadamente un 50% de los datos presentan alguna distorsión o contemplan tokens que son considerados irrelevantes para ser aplicados en los proceso de clasificación. Por ejemplo, signos de puntuación.

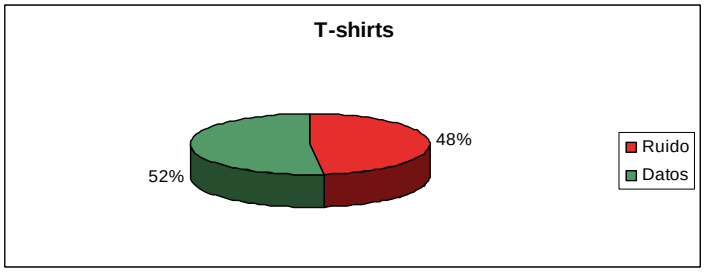

Figura 25: Nivel de ruido en T-Shirts

Es importante aclarar que las causas del ruido son variadas ya que el ingreso de caracteres irrelevantes puede estar asociado a la mala digitación por parte de los agentes de Aduana, pero la fragmentación y fusión de las palabras podría estar relacionada al tratamiento que reciben los datos por parte de los sistemas de Aduana.

Una vez identificadas las distorsiones que presentan los datos se puede cuantificar sus ocurrencias para determinar su distribución en base a las categorías que surgen de la observación de las manifestaciones de ruido más recurrentes, tal como se indica en la Tabla 15.

Tipo	Descripción	Porcentaje
1	Caracteres repetidos	42
2	Símbolos y signos de puntuación	4
3	Palabras fragmentadas	1
4	Palabras fusionadas y errores ortográficos	1

Tabla 15: Ruido presente en registros T-Shirts

La interpretación de los resultados obtenidos indica que para esta partida arancelaria la mayor cantidad de ruido se presenta en la reiterada aparición de caracteres repetidos en las descripciones; esto se encuentra relacionado a la tendencia de algunos agentes de rellenar con secuencias de caracteres del tipo "XXX" o "…" gran parte de las descripciones cuando no conocen las características del producto o simplemente no contemplan realizar una descripción acuciosa de estos.

La importancia de este análisis radica en la posibilidad de disponer de una clasificación sobre las manifestaciones de ruido más frecuente, para que en base a estos resultados se puedan definir mecanismos para atenuarlas.

5.4. MINERÍA DE DATOS

En la etapa de minería de datos, se cuenta sólo con información estructurada donde ahora es posible encontrar pares atributo – valor, recuperados de las descripciones de los productos importados.

El procesamiento se realiza por medio del diseño de un *workflow* realizado en Clementine, el cual permite realizar las siguientes tareas sobre los datos:

- Inspecciones por medio de tablas.
- Selección de atributos.

- Representaciones gráficas.
- Agrupamientos en base a K-means.
- Ordenamiento de datos.
- Generación de reportes.
- Transformaciones de datos.

A continuación se presenta un análisis sobre las importaciones realizado a nivel macro, es decir a nivel de partida arancelaria, donde sólo se conoce que las transacciones pertenecen a la partida arancelaria que se está analizando y un segundo análisis a nivel detallado, donde es posible identificar los elementos que componen una partida.

5.4.1. Análisis a Nivel de Partidas

En las primeras etapas del análisis de los datos estructurados, es posible inspeccionar las entradas por medio de la tabulación de los datos. Estas inspecciones en base a tablas son de utilidad para comprobar que el proceso de carga se está realizando correctamente, y al mismo tiempo validar el correcto funcionamiento de los filtros y transformaciones sobre los datos.

Otra herramienta importante incorporada dentro del *workflow* de Clementine, corresponde a la utilización de histogramas para realizar análisis sobre los distintos campos de los registros que se están analizando. La Figura 26 muestra los resultados obtenidos al realizar un análisis de frecuencia sobre los agentes y las importaciones que estos realizan. El gráfico muestra la distribución de frecuencias ordenada en forma decreciente.

Value	Proportion	%
C29		15,39
I43		9,31
C47		7,06
C12		5,49
C87		5,28
C63		5,27
C48		3,13
Z03		2,9
I64		2,15
F11		2,03
I41		1,82
A33		1,81
C82		1,47
I53		1,47
C11		1,42
F03		1,31
C89		1,25
C14		1,14
I51		1,12
A29		1,03
A30		1,03

Figura 26: Frecuencia de importaciones por agente

Los datos mostrados por el histograma de la figura anterior suele ser de gran ayuda cuando se pretende identificar los agentes en los cuales se concentran los mayores niveles de transacciones.

Después de utilizar el algoritmo K-means, se pueden obtener agrupaciones considerando la cantidad de mercancías (CANT_MERC) versus el valor CIF, que corresponde al valor total de las mercancías, el que se almacena en el campo CIF_ITEM.

Los resultados encontrados se exponen en la Figura 27, donde es posible apreciar un comportamiento conocido; dado por una correlación positiva entre cantidad y valor total. Además, se distingue que un gran volumen de las transacciones se encuentra bajo las 20.000 unidades. La gráfica permite tener una visión general, distinguiendo algunos comportamientos predecibles que tendrían que presentar los datos y al mismo tiempo se logran identificar algunos umbrales para continuar con el análisis. Por otro lado, dentro de los agrupamientos obtenidos llama la atención el cluster-2, donde sus elementos parecen tener los valores más altos para cantidades muy pequeñas de mercancías, lo que sin lugar a duda corresponde a algún tipo de anormalidad.

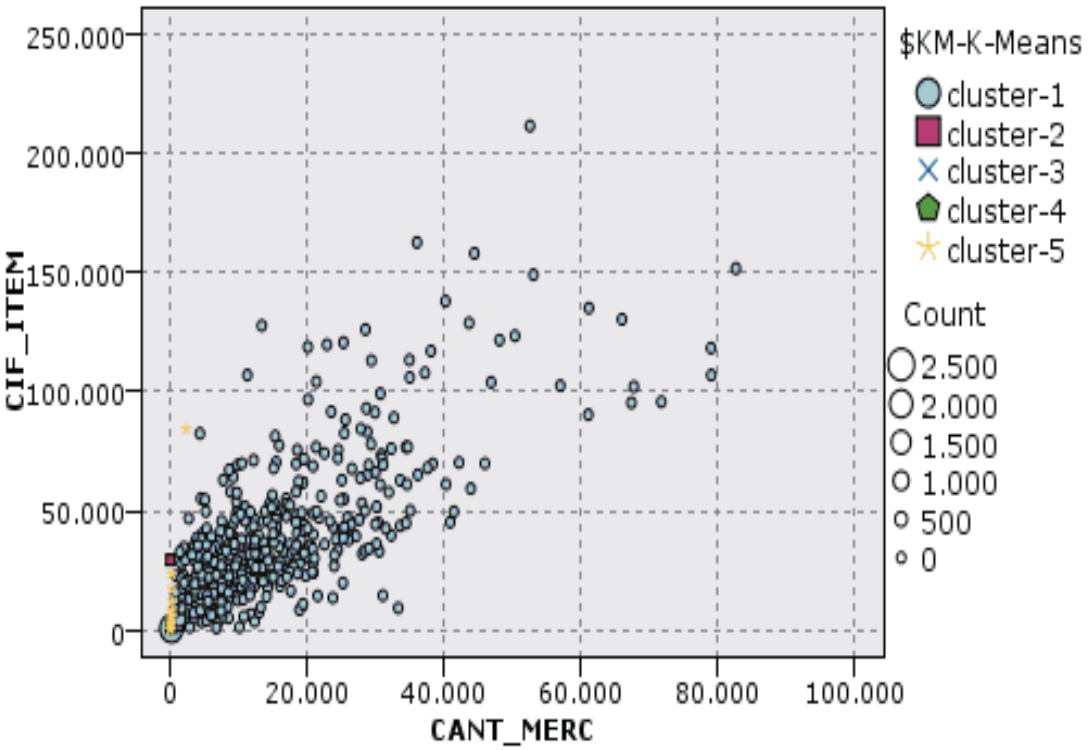

Figura 27: Cluster mercancía versus valor total para T-Shirts

Como en la gráfica de la Figura 27 no es posible identificar claramente las transacciones que presentan comportamientos anómalos, se puede aplicar un cambio de variables para obtener otra vista, en pos de facilitar la identificación de *outliers*. En la gráfica de la Figura 28 se generan agrupamientos considerando el valor unitario que tendrían que tener los productos.

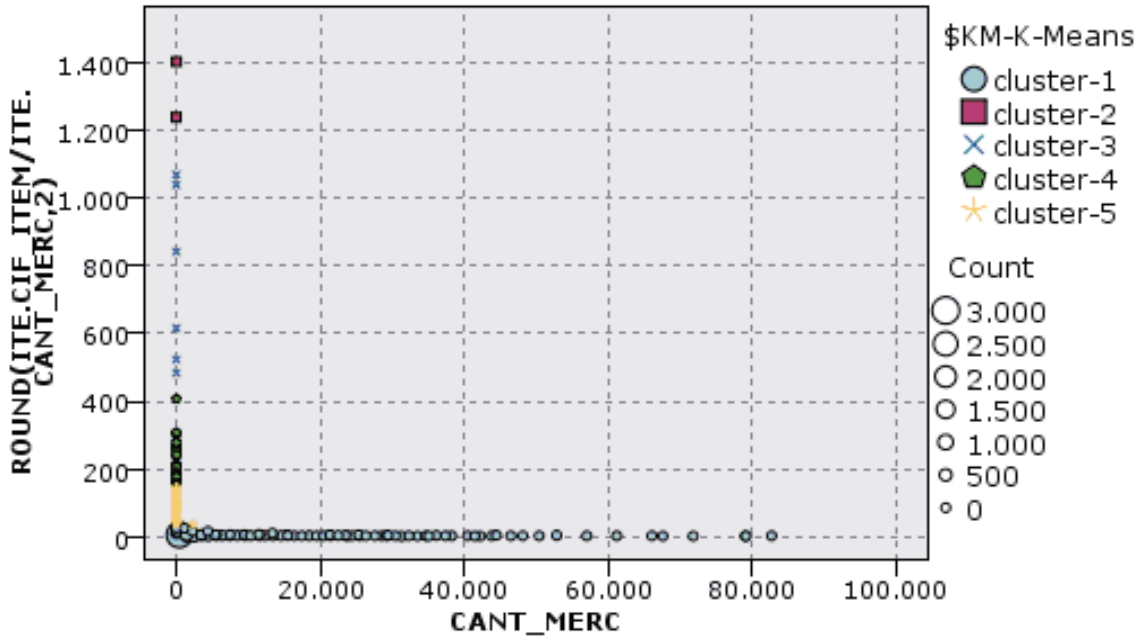

Figura 28: Cantidad de mercancía versus valor unitario- T-Shirts

Los resultados obtenidos por el nuevo agrupamiento son más reveladores, en éstos se aprecia claramente que a mayor cantidad de productos los valores unitarios tienden a disminuir levemente en el cluster-1, el cual contiene a los datos que en su mayoría se consideran normales. Lo anterior contrasta fuertemente con los valores que presentan los cluster restantes, donde se aprecian sobrevaloraciones de productos.

En este punto del análisis, se puede ver que los valores unitarios que presentan las transacciones pueden ser demasiados extremos. A tal punto, que al generar agrupamientos con k=5, se obtienen algunas segmentaciones cuyos contenidos sólo presentan valores extremos. Por otro lado, en una segunda iteración, se requiere focalizarse sobre los elementos del agrupamiento 1 para proceder con la detección de outliers sobre un conjunto de datos que presenta características más homogéneas.

5.4.2. Análisis a Nivel de Productos

En la sección anterior se pudo apreciar un análisis a nivel de partida arancelaria, donde se contemplan campos que pueden ser recuperados directamente desde los sistemas del SNA, pero que aún no cuentan con la información de los productos obtenida por la etapa de minería de texto. Se puede decir entonces, que en el análisis de la sección anterior, la innovación radica en la utilización de técnicas de minería de datos, como el agrupamiento, para organizar los elementos con la finalidad de detectar las anomalías que puedan aflorar productos de la segmentación de éstos.

Además, se pudo observar que es posible identificar anomalías a nivel de los elementos que forman parte de una partida arancelaria, sin incurrir en la necesidad de conocer que tipo de productos conformaban la partida. Esto

último, por las características notoriamente extremas que presentaban los datos respecto del resto. También se demostró que por medio de la utilización de agrupamientos se podían realizar inspecciones sucesivas para identificar las anomalías presentes en los datos.

En esta oportunidad el enfoque será distinto, ya que se incorporará la información de los productos dentro del análisis, para identificar las fluctuaciones de los precios unitarios sobre un producto en particular. A continuación se presenta un breve análisis, donde se utilizan el agrupamiento producido por el algoritmo K-mean en conjunto con el algoritmo LOF.

Lo primero a realizar será seleccionar un producto que tenga una alta frecuencia, para contar con la mayor cantidad de datos posibles de ser incorporados en el análisis. Recurriendo al histograma de la Figura 29, se selecciona la marca "Polo" .

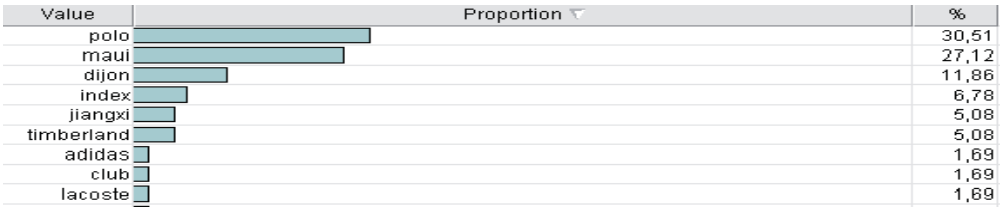

Figura 29: Frecuencia de marcas dentro de la partida T-Shirts

A continuación se procede a identificar todas las transacciones que contienen la marca "Polo". Ahora es posible realizar agrupamientos en base a los atributos de los productos que fueron incorporados como información adicional a cada transacción, gracias a los resultados obtenidos en la etapa de minería de texto. En esta oportunidad los agrupamientos que se obtengan utilizando K-means, identificaran a aquellos elementos que presenten atributos semejantes, lo que permitirá descubrir los distintos productos que se encuentran bajo una marca.

El gráfico de la Figura 30 muestra los resultados para un conjunto reducido de datos, donde se pueden apreciar los valores unitarios que presentan dos de los productos que forman parte de la marca "Polo". Además, en el ejemplo se puede observar que para la marca analizada existen sólo dos productos clasificados, donde el que tiene mayores ocurrencias se encuentra en el agrupamiento 1.

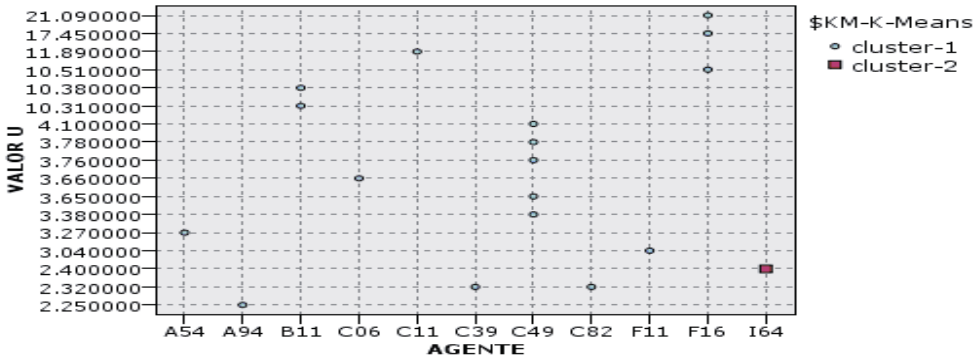

Figura 30: Precio unitario por producto y agente

Si es posible identificar los productos por medio de agrupamiento de muestras con atributos similares, se cuenta con una potente herramienta para realizar un análisis de las fluctuaciones de los precios unitarios para un producto en particular. En este punto es posible realizar nuevamente agrupamientos por medio de K-means para identificar aquellos valores que más se apartan de la media. Pero en esta oportunidad el análisis se realizará utilizaran los resultados obtenidos por el algoritmo LOF (con k=10), el cual a partir de la densidad de los conjuntos de datos, identificará aquellas regiones con menor densidad, las que tienden a presentar *outliers*. La metodología que se aplica en este punto del análisis es muy similar a lo expuesto en [98], donde si se logra contar con la información que permita individualizar un producto, entonces es posible utilizar sólo la dimensión del valor unitario para proceder con un análisis de anomalías. Lo anterior facilita notablemente el procedimiento y en este caso esa información esta disponible gracias a la recuperación de los atributos desde las descripciones.

La Figura 31 muestra las puntuaciones generada por LOF, donde aquellos elementos que se encuentran circunscritos, presentan una alta puntuación, la que es directamente proporcional al grado anómalo que presenta el objeto.

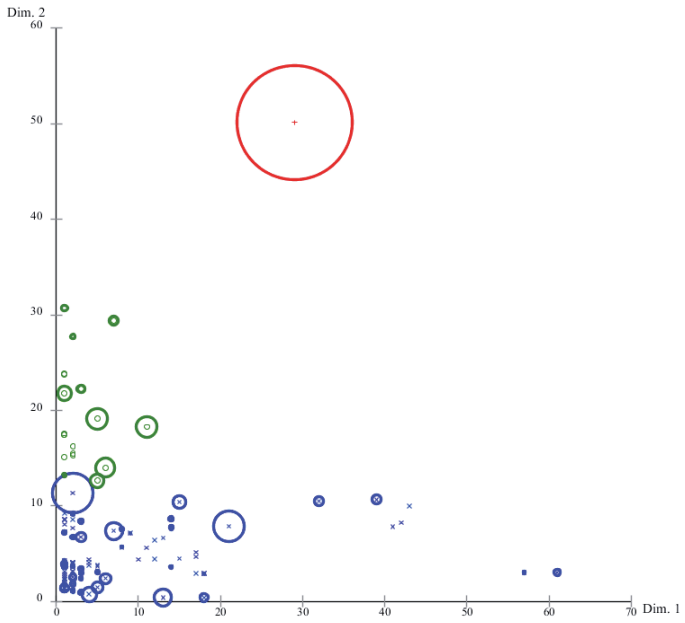

Figura 31: Identificación de outliers usando LOF

Una vez que se ha obtenido la puntuación LOF para cada precio unitario de las transacciones de un producto especifico, el paso siguiente es normalizar el factor obtenido en una escala de 0 a 1. Para lo anterior se puede utilizar una función de normalización como *SoftMax*. Una vez normalizadas las puntuaciones para cada transacción, es posible obtener el ranking de los elementos anómalos que serán seleccionados para ser inspeccionados.

En esta etapa final existen varias enfoques que pueden ser utilizados para seleccionar a los elementos sobresalientes del ranking, (1) es posible seleccionar a aquellos elementos que sobrepasen un límite definido, como es el caso utilizado en el análisis de K-mean, donde se seleccionaron aquellos objetos que estaban por sobre o bajo el doble de la media , o (2) utilizar la puntuación LOF normalizada, para seleccionar las N primeras puntuaciones que presenten más de un 60% de probabilidad de ser anómalas.

Una vez obtenidos el conjunto de elementos anómalos seleccionados, estos son entregados a un experto para que realice una segunda selección, donde a los resultados obtenidos en esta última selección se les aplicará un proceso de fiscalización.

5.5. CONCLUSIONES

Por tratarse de una investigación donde se trabaja con transacciones reales, resulta complicado realizar una evaluación sobre los resultados obtenidos, principalmente porque es necesario aplicar un proceso de fiscalización sobre las mercancías, para de esta forma confirmar la presencia de fraude, lo que obliga a establecer una coordinación con el departamento de fiscalización de Aduana para llevar a cabo un proceso que suele ser costoso. Como el escenario anterior es demasiado complejo para llevarlo a la práctica, una solución razonable es trabajar con transacciones históricas, donde se podrá contar con los resultados obtenidos producto de fiscalizaciones que ya fueron aplicados en años pasados. De esta forma, la evaluación contempla la selección de un conjunto de datos con alta probabilidad de fraude, obtenidos por medio de la salida que entrega la solución diseñada, sobre este conjunto de datos Aduana puede consultar si alguna de las transacciones seleccionadas fue fiscalizada y posteriormente comparar los resultados obtenidos para ver si presenta fraude por subvaloración. Teniendo presente lo anterior, se seleccionaron 300 registros que presentaban una alta distorsión en los valores unitarios de distintos productos. Estos fueron enviados al Servicio Nacional de Aduana para que cruzara esta información, con los resultados de las fiscalizaciones realizadas durante el año 2010. Los datos entregados por el SNA indican que de las muestras que presentaban una fiscalización durante el año 2010, un 40% contemplaba una denuncia asociada a fraude por subvaloración, donde también se encontraron algunos casos de contrabando.

Sobre los casos clasificados como sobrevaloración, se aclaró que uno de estos se encontraba claramente justificado ya que se trataba de la importación de camisetas deportivas para coleccionistas, de ahí el alto valor unitario que presentaban los productos. Los casos restantes fueron entregados a expertos de fiscalización para analizar las causas de ese comportamiento. Sobre lo anterior, puede resultar extraño evaluar un posible fraude cuando se trata de sobrevaloración, pero la razón de fondo radica en que dichas transacciones tiene que estar vinculada a una adulteración, intencional o involuntaria, de los campos cantidad o valor debido a lo extremo de éstos con respecto a los valores más frecuentes. Por esto, no se descarta la posibilidad de que se obtenga un valor unitario elevado como producto de un registro inadecuado de las cantidades reales que se están importando.

Respecto de la cantidad de productos que se pueden recuperar de los datos, ésta se mantiene en el orden de un 50%. Las causas de esto radican principalmente, en que un producto de software de estas características necesita un

periodo de tiempo mayor, para alcanzar un rendimiento acorde a un entorno de producción. Otro factor importante radica en que una gran cantidad de transacciones no presentan los atributos mínimos para lograr una clasificación y en el peor de los casos no cuentan con una cantidad de tokens aceptables para superar los filtros de preevaluación de datos. Para remediar esta última limitante, se hace necesario crear componentes que permitan validar el ingreso de las descripciones por medio de componentes que puedan identificar los componentes mínimos deseables en una descripción por medio de una evaluación semántica.

Respecto de la identificación de *outliers* por medio de técnicas no supervisadas, es posible optimizar el proceso recurriendo a la utilización conjunta de más de un algoritmo. Pero por la complejidad inherente a un proceso de estas características donde se tienen que evaluar y optimizar el trabajo conjunto de distintas técnicas, se ha considerado plantear este desafío como parte de un trabajo futuro.

6. DISCUSIÓN Y CONCLUSIONES

A modo general, desde el punto de vista del trabajo realizado, se puede concluir que se ha logrado diseñar un *framework* (compuesto por componentes de código abierto y cerrado) cuyo objetivo principal es transformarse en una herramienta de apoyo para el área de inteligencia aduanera. La solución obtenida permite recuperar información desde los atributos que contienen la descripción de las mercancías, información que se encontraba disponible pero que no había podido ser utilizada en un proceso automatizado, ya que el SNA no contaba con las herramientas necesarias para trabajar con texto no estructurado. Dicha información ahora puede ser utilizada para identificar posibles fraudes de subvaloración de productos, por medio de la aplicación de minería de datos, que permite la inspección de grandes volúmenes de datos, donde se pueden incorporar múltiples variables relevantes al análisis (países de origen, puertos, agentes de Aduana, etc.). Además, ahora el análisis se puede extender para considerar los productos como unidad de trabajo, lo cual facilita la detección de anomalías asociadas a la subvaloración de estos. La incorporación de este nuevo paradigma ha demostrado ser efectivo, alcanzando una predicción de un 40% en base a la información disponible sobre los resultados de las fiscalizaciones realizadas el año 2010.

Sobre la investigación realizada, se tiene la seguridad de que existe una gran carencia de publicaciones que enfrenten los problemas de fraude desde el punto de vista de las operaciones de comercio exterior. Lo mismo ocurre con la extracción de productos a partir de sus descripciones. Lo que es peor, gran parte de éstos se limitan sólo a enfrentar áreas de aplicación muy explotadas, las cuales gozan de popularidad en entornos académicos, tal es el caso de la determinación del grado de satisfacción de los clientes, trabajos que están claramente orientados a cubrir las necesidades de las empresas de *retail* y *marketing*.

En lo que respecta a la incorporación de tecnología de la información (TI), se conoce que ésta por si sola no entrega valor al negocio. La incorporación de la TI tiene que estar alineada con las necesidades del negocio para que se transforme en un aporte y no en un gasto innecesario. Es por éstas razones que el estudio se centra en cubrir necesidades consideradas importantes por el área de inteligencia, pero que no podían ser resueltas de manera apropiada, con la tecnología disponible. De ésta forma, el presente trabajo no sólo entrega una propuesta tecnológica con un enfoque innovador, sino que también puede ser considerado como un proyecto estratégico, que permita interiorizar a las jefaturas y personal técnico involucrado, sobre las herramientas disponibles para potenciar partes claves del actual proceso de fiscalización.

En la tesis no sólo se contempla la incorporación de nuevas tecnologías integradas en una solución innovadora, sino que también se propone un cambio en la forma en que se realizan los análisis. Este cambio de paradigma propone profundizar el análisis a nivel de productos, por medio del descubrimiento de los elementos que componen una partida arancelaria. Esto se traduce en la incorporación de los productos como partes de una nueva vista, donde se entrega al analista información relevante, la que se considera fundamental para ser utilizada en un análisis de precios para ser aplicado a la detección de subvaloraciones. Dicho cambio, afecta directamente a la cobertura a la que puede aspirar el proceso de análisis, permite además facilitar la identificación de fluctuaciones relevantes en el precio de aquellos productos que forman parte de un grupo homogéneo, lo que facilitar la identificación de patrones de

comportamiento, que tienden a permanecen ocultos al trabajar con agrupaciones genéricas, como es el caso de trabajar sólo a nivel de partida arancelaria. En este último punto, se tendrá que considerar un periodo de tiempo mayor para comprobar si este cambio de paradigma cumple con las expectativas planteadas.

Otro aspecto importante a considerar son las conclusiones asociadas al *framework* como un producto de software, las cuales se enmarcan dentro de las apreciaciones técnicas, que surgen en base a la experiencia obtenida durante su desarrollo.

- La calidad de los datos es un requisito fundamental para que el *framework* pueda realizar el proceso de clasificación. A pesar de que se logra atenuar parte del ruido presente en las descripciones, se descubrió que en algunos casos, la información provista en éstas es insuficiente para permitir una clasificación. Por esta razón, se recomienda evaluar la posibilidad de no sólo validar que el campo descriptivo contenga datos, sino que también se considere utilizar un filtro, que apoyado por técnicas de procesamiento de lenguaje, determine si el contenido de la descripción cuenta con los atributos mínimos necesarios para ser aceptada. Es importante tener presente que dicho filtro tiene que poder identificar la presencia de una cantidad mínima de atributos sin que esto entorpezca notoriamente la agilidad del actual proceso de validaciones.

- Los mecanismos utilizados para la atenuación de ruido demostraron ser eficientes. Sin embargo es posible mejorar las correcciones ortográficas que se apoyan en la utilización de un modelo de lenguaje. Para cumplir con esto se propone utilizar el procesamiento de N-gramas a nivel de letras, apoyado con la incorporación de manejo de co-ocurrencias en los tokens más significativos.

- La disponibilidad de recursos lingüísticos para el español es muy limitada y los corpus de entrenamiento son muy costosos. Por estas razones, el análisis de texto no estructurado se basa principalmente en un enfoque estadístico a nivel muy experimental, donde las palabras y las relaciones a nivel de N-gramas cumplen un rol fundamental. Este enfoque demostró ser efectivo, y al mismo tiempo le entrega la flexibilidad necesaria al *framework* para reconocer cualquier producto nuevo por medio de la extracción de características relevantes.

- La utilización de diccionarios puede reducir notablemente la dimensionalidad de los datos, lo que favorece el análisis y la precisión en los clasificadores. Considerando que el vocabulario presente en las descripciones es bastante limitado, resulta apropiado contemplar la creación de diccionarios de contexto, para distintos tipos de mercancías, sin que esto implique un costo significativo.

Por otro lado es necesario reconocer, que este trabajo no hubiese podido ser llevado a cabo, si no fuera por los conocimientos obtenidos durante el programa de magíster. En este último punto las asignaturas vinculadas con inteligencia de negocio cumplieron un papel fundamental, ya que pasaron a constituir la base teórica de este trabajo y al mismo tiempo sirvieron de incentivo para extender los conocimientos adquiridos a otras áreas de aplicación como el manejo de textos no estructurados.

El aporte de este trabajo a la formación profesional del autor es indiscutible. En los contenidos abordados se seleccionaron áreas de TI que gozan de gran demanda en la actualidad; tal es el caso de la minería de texto la cual es una herramienta que ha adquirido gran relevancia debido a la explosión de contenidos en Internet, la masificación de las redes sociales y el aporte que realiza dentro de las empresas para recuperar conocimiento a partir de fuentes no estructuradas, las cuales según las estadísticas corresponde al 80%[3] de los datos relevantes de una empresa. Además se trató de incorporar los últimos avances que existen en la materia, esto se puede apreciar claramente en las referencias seleccionadas. Por último, existe la iniciativa de utilizar los conocimientos adquiridos por medio de este trabajo en el desarrollo de proyecto asociados al análisis de sentimientos, la administración documental, detección de fraudes y en análisis de seguridad informática, utilizando minería gráfica.

[3] Fuente Gartner

7. TRABAJO FUTURO

El *framework* obtenido corresponde a una solución que por su característica de aprendizaje (supervisado), para la recuperación de pares atributo–valor, puede ser mejorada a una propuesta que implique un menor costo en el entrenamiento. Sobre lo anterior, es recomendable analizar el trabajo realizado por Blum [73] sobre la posibilidad de utilizar datos etiquetados y sin etiquetar junto a algoritmos semi-supervisados, técnica que recibe el nombre de co-entrenamiento. Dicho trabajo resulta ser apropiado en aquellos entornos donde la cantidad de datos a considerar es muy extensa y se dispone de un pequeño conjunto de datos etiquetados, los cuales son utilizados como datos semilla. El gran aporte de esta técnica radica en extender la clasificación a aquellos datos que no habían sido considerados en el entrenamiento supervisado, tal como lo demuestra Rayid Ghani y Katharina Probst junto a sus grupos de trabajo, quienes logran realizar clasificaciones de productos exitosas utilizando co-entrenamiento.

Por otro lado, es posible mejorar las técnicas que son utilizadas en la minería de texto, incorporando algoritmos que mejoren el procesamiento del lenguaje y que se especialicen en un vocabulario asociado al contexto. Dentro de esto último, se puede destacar la mejora del modelo de lenguaje creado para realizar las correcciones ortográficas y la supresión del ruido.

Otro aspecto importante del cual se puede sacar un mayor provecho, corresponde a la generación grafos para facilitar el análisis por medio de la visualización de vinculaciones. Actualmente el *framework* sólo utiliza esta característica para analizar los datos que componen las descripciones, entregando una red de nodos que representa la frecuencia de las palabras y el grado de relación entre estas. Lo anterior se puede extender a un escenario más complejo, donde el análisis se puede centrar en las transacciones de comercio exterior y las relaciones que existen entre las distintas entidades que forman parte en este proceso. De esta forma, a la luz de un modelamiento apropiado, es posible obtener conocimiento sobre comportamientos que no pueden ser identificados por herramientas que no contemplan las relaciones que surgen entre los nodos. Algunos casos de éxito de análisis de vinculaciones, donde destacan situaciones asociadas al lavado de dinero, narcotráfico y casos de fraude, pueden ser consultados en [41-48].

En lo referente a algoritmos para la detección de *outliers* es posible mejorar el enfoque adoptado por medio de la utilización de múltiples algoritmos que se complemente, tal como es propuesto y justificado en [55]. Además, las mejoras en el rendimiento de los algoritmos no supervisados también pasa por identificación los valores óptimos para variables sensibles como el tamaño del vecindario para LOF y la cantidad de cluster para K-means, algo que quedó fuera del alcance de este trabajo.

Para terminar, es importante recalcar, que a futuro se pretende obtener un producto completamente desarrollado en software de tipo *Open S*ource, por medio de la migración de las funcionalidades de agrupamiento provistas por la herramienta Clementine a su equivalente en código abierto. Lo anterior tiene sentido, si se considera que muchas de

las versiones actuales de productos *Open S*ource para minería, presentan una cantidad mayor de técnicas a las que disponibles en Clementine 12[4].

[4] SPSS fue adquirido el año 2009 por IBM

ANEXO A: HERRAMIENTAS PARA MINERÍA DE TEXTO

A continuación se presenta un breve listado con productos tanto pagados como libres, que pueden ser utilizados en tareas de minería de texto.

Autonomy

http://www.autonomy.com/

Autonomy es una empresa líder a nivel mundial en el sector de software de infraestructura. Ofrece un conjunto de productos para ser utilizados en desarrollos dentro de las organizaciones, en intranets o Internet. Además de poseer una robusta infraestructura para el procesamiento de lenguaje natural, cuanta también con funcionalidades avanzadas que le permite procesar audio y video.

NLTK

http://www.nltk.org/

Natural Lenguage Toolkit, corresponde a una API desarrollada en Python por la Universidad de Melbourne bajo licencia GNU, para ser utilizada en tareas de procesamiento de lenguaje natural. NLTK contiene módulos para desarrollar etiquetadotes, *chunking*, *parsers* completos, preprocesamiento y agrupamiento.

OpenNLP

http://incubator.apache.org/opennlp/

Corresponde a un conjunto de proyectos heterogéneos asociados con el procesamiento de lenguaje natural utilizando lenguaje Java. Dentro de las funcionalidades disponibles se encuentran; tokenización, segmentación de sentencias, etiquetadores POS, extracción de entidades, *chunking*, *parsers* y resolución de correferencias.

Stanford NLP software

http://nlp.stanford.edu/software/index.shtml

Corresponde a un conjunto de herramientas publicadas por la Universidad de Stanford desarrolladas en Java con licencia GNU. Las herramientas disponibles contemplan algoritmos etiquetadores POS, clasificación de texto y *parsers PCFG.*

Freeling

http://nlp.lsi.upc.edu/freeling/

Es un set de herramientas desarrolladas en C++ por la universidad politécnica de Catalunya bajo licencia GNU. Dentro de las funcionalidades contempladas se encuentra; detección de sentencias, análisis morfológico, reconocimiento de entidades, etiquetadotes POS, shallow parsers, parser de dependencias.

Gate

http://gate.ac.uk

General Architecture for Text Engineering (GATE), corresponde a una suite de herramientas desarrolladas en Java por la universidad de Sheffield en 1995. Gate está provisto de un IDE que permite contar con un entorno de desarrollo integrado para realizar tareas de procesamiento de lenguaje natural y extracción de información.

SAS Text Analytic

http://www.sas.com/text-analytics/index.html

SAS Text Analytic incorpora avanzadas capacidades lingüísticas dentro de su *core* para minería de datos llamada SAS Enterprise Miner. Text Analytic se encuentra dividido en los siguientes componentes:

- Categorizador de contenido empresarial:
- Análisis de sentimientos
- Administrador de ontologías
- Minería de texto

ANEXO B: HERRAMIENTAS PARA MINERÍA DE DATOS

A continuación se presenta un breve listado con productos tanto pagados como libres, que pueden ser utilizados en tareas de minería de datos.

Orange

http://orange.biolab.si/

Es un entorno de trabajo desarrollado en C++, Python y QT, para implementar flujos de trabajo con tareas de minería de datos y aprendizaje automatizado, basado en componentes. Dentro de sus características se destaca la amistosidad de su interfaz gráfica, que permite una rápida y versátil programación. El producto contiene un completo set de componentes para el preprocesamiento, selección de características, modelamientos, evaluación de modelos y técnicas de exploración de datos.

RapidMiner

http://rapidminer.com/

Es un entorno para minería de datos y aprendizaje automatizado que ha sido diseñado para entregar soluciones en un ambiente de producción. La herramienta utiliza su interfaz gráfica para definir flujos de trabajos los cuales quedan almacenados en estructuras de archivos XML. Rapid Miner dispone de más de 500 operaciones para ser utilizadas en aprendizaje automatizado y además permite integrarse con Weka.

Weka

http://www.cs.waikato.ac.nz/~ml/weka/

Weka (Waikato Environment for Knowledge Analysis), es una suite escrita en Java que permite efectuar tareas típicas de minería de datos como; preprocesamiento, agrupamiento, clasificación regresión, visualización y selección de características. El entorno también permite acceso a bases de datos SQL utilizando el soporte de conectividad que entrega Java.

KNIME

http://www.knime.org

KNIME (Konstanz Information Miner), es una plataforma amistosa, que permite el procesamiento, análisis, integración y exploración de los datos. Permite al usuario definir flujos de datos o *pipelines* donde se puede ejecutar selectivamente algunos o todos los pasos de análisis para posteriormente estudiar los resultados, modelos y vistas interactivas que se generan. El entorno se encuentra desarrollado en Java sobre Eclipse lo que le permite soportar la incorporación de *plugins* para obtener funcionalidades adicionales para texto, imágenes, series de tiempo y la integración con otros proyectos open sources como R, Weka, Chemestry y LibSVM.

Oracle Data Mining

http://www.oracle.com/technetwork/database/options/odm/index.html

Oracle Data Mining (ODM), se encuentra implementada en el kernel de la Base de datos Oracle 11*g* Enterprise Edition. El producto permite producir información predictiva y crear aplicaciones con inteligencia de negocio integrada. La funcionalidad de la extracción inteligente de datos incorporada en Oracle Database 11*g*, faculta a los usuarios para realizar operaciones de búsqueda de patrones y recuperación de conocimientos ocultos en sus datos.

REFERENCIAS

[1] Raúl Monge, Juan Francisco Germaín, Eduardo Valenzuela, John Atkinson, Carlos Rosales y Cristián Neumann, "Desarrollo de un Prototipo de Fiscalización Inteligente en el Marco del Rediseño de la Fiscalización del Servicio Nacional de Aduanas de Chile", Informe Final de Consultoría, U.T.F.S.M., 30 Junio 1998.

[2] Propuesta de Rediseño de la Fiscalización, Servicio Nacional de Aduanas, 2000.

[3] El Proceso de Fiscalización de Aduana, Subdirección de Fiscalización, Servicio Nacional de Aduanas, Septiembre 2000.

[4] Consultoría "Viabilidad técnica de redes neuronales para el apoyo en línea de la fiscalización" Kyber S.A., Octubre 2000.

[5] Consultoría "Modelo de importación DIN" generado por Computer Associates en conjunto con el grupo de especialistas del departamento de Inteligencia Aduanera, Octubre 2006.

[6] Inteligencia y gestión de riesgos en la fiscalización de acuerdos internacionales de medio ambiente: La experiencia de la Aduana Chilena. Servicio Nacional de Aduanas, 2005.

[7] Christopher D. Manning, Hinrich Schütze. Foundations of statistical natural language processing. The MIT press, 2000.

[8] Graham Wilcock. Introduction to linguistic annotation and text analytics. Morgan & Claypool Publishers, 2009.

[9] Claude E. Shannon . A mathematical theory of Communications. Bell System Technical Journal, 1948.

[10] Claude Shanon. Communication in the presence of noise. Proceedings of the IRE. 37(1):10-21. 1949

[11] Wiener, Norbert. Cybernetics, or Control and Communication in the Animal and the Machine. Paris: Hermann and Co., Cambridge, MA: The Technology Press, and New York, NY: John Wiley and Sons, 1948

[12] Wiener, Norbert. The Human Use of Human Beings: Cybernetics and Society. Boston, MA: Houghton Mifflin, 1950.

[13] M. A. Martí y J. Llisterri Tratamiento del lenguaje natural. Edicion universitaria, 2002

[14] Daniel Jurafsky and James H. Martin. Speech and language processing, Prentice Hall, 1999.

[15] Jesús Vilares Ferro. Aplicaciones del procesamiento del lenguaje natural en la recuperación de información en español. Tesis doctoral, Universidad de la coruña, Mayo del 2005.

[16] M. A. Martí y J. LListerri. Tratamiento del lenguaje natural", edición universitaria de Barcelona 2002.

[17] Ronen Feldman y James Sanger. "The text Mining handbook", Cambridge University Press 2007.

[18] Anne Kao and Stephen R. Poteet. "Natural language processing and text mining", Springer, 2007.

[19] A. Kloptchenko, T. Eklun, B. Back , J. Karlson, H. Vanharanta and A. Visa. "Combining data and text mining techniques for analyzing financial reports, Eighth Americas conference on information systems, 2002.

[20] U. Y. Nahm and R. J.Mooney. Textmining with infor-mation extraction. In AAAI 2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases, 2002.

[21] Martin F. Porter. An algorithm for suffix stripping. Program, 14(3):130{137, 1980

[22] Julie B. Lovins. Development of a stemming algorithm. Mechanical Translation and Computational Linguistics, 11(1{2):22 {31, 1968

[23] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. "Introduction to information retrieval", Cambridge University Press 2008.

[24] Michael W. Berry. "Survey of text Mining: Clustering, classification, and retrieval", Springer 2003.

[25] Ricardo Baeza-Yates and Berthier Ribeiro-Neto. Modern Information Retrieval. Addison Wesley and ACM Press, Harlow, England, 1999.

[26] Melvin Maron y John L. Kuhns. "On relevance, probabilistic indexing, and information retrieval", Journal of the ACM 7(3):1216-244, 1960.

[27] William B. Frakes and Ricardo Baeza-Yates. Information Retrieval Data Structures & Algorithms

[28] Gerardo Canfora, Luigi Cerulo. A taxonomy of information retrieval models and tools. Journal of computing and information technology CIT 12, 2004.

[29] Amit Singhal. Modern Information Retrieval: A brief overview. IEEE Data Engineering Bulletin, 2001.

[30] Marie Francine Moens. "Information Extraction: Algorithms and prospects in a retrieval context", Springer 2003.

[31] Damerau, Fred J. 1964. A technique for computer detection and correction of spelling errors, Communications of the ACM 7(3):171-176.

[32] Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady.

[33] Kleene, Stephen C. (1956). "Representation of Events in Nerve Nets and Finite Automata". In Shannon, Claude E.; McCarthy, John. Automata Studies. Princeton University Press. pp. 3–42
[34] David L. Olson and Dursun Delen, "Advanced data mining techniques", Springer, 2008
[35] Clifton Phua, Vincent Lee, Kate Smith and Ross Gayler. A Comprehensive Survey of Data Mining-based Fraud Detection Research.
[36] Da Ruan, Guoquing Chen, Etienne E. Kerre and Geert Wets. "Intelligent data mining, techniques and applications", Springer, 2005
[37] Mehmed Kantardzic "Data mining concepts, models, methods and algorithms"
[38] Yury Yineth Roa, "Minería de textos médicos web, un enfoque hacia la clasificación en categorías"
[39] A. McCallum and K. Nigam. A comparison of event models for naive Bayes text classication. In Learning for Text Categorization: AAAI Workshop, 1998
[40] Brockett P, Xiao X and Derris R. "Using Kohonen Self Organizing Feature Map to Uncover Automobile Bodily Injury Claims Fraud", Journal of Risk and Insurance, USA, 1998.
[41] Jesús Mena. Investigative Data Mining for Security and Criminal Detection, Butterworth Heinemann 2003.
[42] He H, Wang J, Graco W y Hawkins S. "Application of Neural Networks to Detection of Medical Fraud", Expert Systems with Applications, 1997.
[43] Chan P, Fan W, Prodromidis A and Stolfo S. "Distributed Data Mining in Credit Card Fraud Detection", IEEE Intelligent Systems, 1999.
[44] Weatherford M. "Mining for Fraud", IEEE Intelligent Systems, July/August Issue, 2002.
[45] Fawcett T and Provost F. "Adaptive fraud detection", Data Mining and Knowledge Discovery, Kluwer, 1997.
[46] Olusola Adeniyi Abidogun, "Data mining, fraud detection and mobile telecommunications: call pattern analysis with unsupervised neuronal network", Master's thesis, university of Western Cape, 2005.
[47] Tom Fawcett and Foster Provost. Adaptive Fraud Detection. Data Mining and Knowledge Discovery, Volume 1, Number 3, 291-316.
[48] Wilfredy Santamaria Ruiz. Técnicas de Minería de Datos Aplicadas en la Detección de Fraude:Estado del Arte.
[49] Efstathios Kirkos and Charalambos Spathis and Yannis Manolopoulos. Data Mining techniques for the detection of fraudulent financial statements. Expert Systems with Applications, Vol 32, pp 995-1003, May 2007.
[50] Tom Fawcett and Foster Provost. Adaptive Fraud Detection. Data Mining and Knowledge Discovery, Volume 1, Number 3, 291-316.
[51] Andrew Gelman, John B. Carlin, Hal S. Stern, Donald B. Rubin. Bayesian Data Analysis. Chapman & Hall/CRC 2004.
[52] Using data mining to detect fraud. SPSS White paper – technical report 2000.
[53] Rosie Jone. Learning to extract entities from labeled and unlabeled text. University of Utah 2005.
[54] Anoop Sharka. Combining labeled and unlabeled data in statistical natural language parsing.University of Pennsylvania 2002.
[55] H. Kuna, G. Pautsch, M. Rey, C. Cuba, A. Rambo, S. Caballero, A. Steinhilber, R. García Martínez y F. Villatoro. Avances en procedimientos de la explotación de información con algoritmos basados en la densidad para la identificación de outliers en bases de datos, Universidad Nacional de Rosario. Rosario. Santa Fe. Argentina.
[56] Fernando Pereira. Beyond word N-grams. In David Yarovsky and Kenneth Church, editors, Proceedings of the Third Workshop on Very Large Corpora, pages 95 – 106, Somerset, New Jersey, 1995.
[57] Clifton Phua, Damminda Alahakoon, and Vincent Lee. Minority Report in Fraud Detection: Classification of Skewed Data. School of Business Systems, Faculty of Information Technology Monash University
[58] Clifton Phua, Vincent Lee, Kate Smith and Ross Gayler. A Comprehensive Survey of Data Mining-based Fraud Detection Research.
[59] W. Frawley, G. Piatetsky-Shapiro y C. Matheus. "Knowledge Discovery in Databases: An Overview". AI Magazin
[60] P. Chapman, J. Clinton, R. Kerber, T. Khabaza, T. Reinartz, C. Shearer and R. Wirth. "CRISP-DM 1.0 step-by-step data mining guide", 2000.
[61] Svetlana Cherednichenko, "Outlier detection in clustering", Master's thesis, university of Joensuu, 2005
[62] Irad Ben-Gal. Outliers Detection. Department of Industrial Engineering, Tel-Aviv University.
[63] M. O. Mansur, Mohd. Noor Md. Sap. Outlier Detection Technique in Data Mining: A Research Perspective. Faculty of Computer Science and Information Systems, Annual Research Seminar 2005.
[64] Jesus Mena. Investigative Data Mining for Security and Criminal Detection, Butterworth Heinemann 2003.
[65] Zengyou He and Xiaofei Xu and Joshua Zhexue Huang and Shengchun Deng. Mining class outliers: concepts, algorithms and applications in CRM. Expert Systems with Applications, Vol 27, pp 681-697, Nov 2004.

[66] Zengyou He and Xiaofei Xu and Shengchun Deng. A Fast Greedy Algorithm for Outlier Mining. Computer Science, 2005

[67] Kaustav Das and Jeff Schneider. Detecting anomalous records in categorical datasets. Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining, pp 220-229, New York 2007.

[68] Tianming Hu and Sam Y. Sung. Detecting pattern-based outliers. Pattern Recognition Letters, Vol 24, pp3059-3068,Dec 2003.

[69] J. A. Fernandez Pierna and F. Wahl and O. E. de Noord and D. L. Massart. Methods for outlier detection in prediction. Chemometrics and Intelligent Laboratory Systems, Vol 63, pp 27-39, Aug 2002.

[70] Lipika Dey and Sk. Mirajul Haque. Opinion mining from noisy text data. Springer-Verlag 2009

[71] Clark, A.: Preprocessing very noise text. In: Proceedings of Workshop on Shallow Processing of Large Corpora, Corpus Lin.

[72] Nasukawa, T., Punjani, D., Roy, S., Subramaniam, L.V., Takeuchi, H.: Adding Sentence Boundaries to Conversational Speech Transcriptions using Noisily Labelled Examples. In: Proceedings of IJCAI-2007 Workshop on Analytics for Noisy Unstructured Text, pp. 71–78, Hyderabad, India (2007).

[73] A. Blum and T. Mitchell. Combining labeled and unlabeled data with co-training. In COLT-98,

[74] Kamal Nigam and Rayid Ghani. Analyzing the Effectiveness and Applicabilityof Co-training. 2000 1998

[75] K. Nigam, A. McCallum, S. Thrun, and T. Mitchell. Text classification from labeled and unlabeled docu-ments using EM. Machine Learning, 39(2/3):103–134, 2000.

[76] R. Ghani and A. E. Fano. Building recommender sys-tems using a knowledge base of product semantics. In Proceedings of the Workshop on Recommendation and Personalization in ECommerce at the 2nd Interna-tional Conference on AdaptiveHypermedia and Adap-tive Web based Systems, 2002.

[77] Katharina Probst, Rayid Ghani, Marko Krema and Andy Fano. Semi-Supervised Learning to Extract Attribute-Value Pairs from Product Descriptions on the Web, 2007.

[78] B. Liu, M. Hu, and J. Cheng. Opinion observer: Analyzing and comparing opinions on the web. In Proceedings of WWW 2005, 2005.

[79] Rayid Ghani and Andrew E. Fano. Using Text Mining to Infer Semantic Attributes for Retail Data Mining. 2002

[80] Ana-Maria Popescu and Oren Etzioni. Extracting Product Features and Opinions from Reviews. 2005.

[81] Rayid Ghani, Katharina Probst, Yan Liu, Marko Krema, Andrew Fano. Text Mining for Product Attribute Extraction. 2006

[82] Steven Bird, Ewan Klein and Edgard Loper. Natural language processing with python. O'Reilly 2009.

[83] Xindong Wu, Vipin Kumar, J. Ross Quinlan, Joydeep Ghosh, Qiang Yang,Hiroshi Motoda, Geoffrey J. McLachlan, Angus Ng, Bing Liu, Philip S. Yu, Zhi-Hua Zhou, Michael Stein. Top 10 algorithms in data mining. Springer 2007.

[84] Eric Brill and Robert C. Moore. An Improved Error Model for Noisy Channel Spelling Correction. Microsoft Research One Microsoft Way, 2000.

[85] Silviu Cucerzan and Eric Brill. Spelling correction as an iterative process that exploits the collective knowledge of web users. Microsoft Research One Microsoft Way.

[86] Kernighan, M., Church, K., Gale, W.: A spelling correction program based on a noisy channel model. In: Proceedings of COLING 1990, pp 205–210 (1990).

[87] Mark Graven. Learning to extract relations from MEDLINE. In papers from the sixteenth national conference on artificial intelligence (AAAI-99) workshop on machine learning for information extraction.

[88] Razvan Bunescu, Raymond Mooney Arun Ramani y Edward Marcotte. Integrating Co-occurrence Statistics with Information Extraction for Robust Retrieval of Protein Interactions from Medline, 2006.

[89] Padraig Cunningham. Dimension reduction. Technical report UCD-CSI, 2007.

[90] P. Ponmuthuramalingam and T. Devi. Effective dimension reduction techniques for text documents. International Journal of Computer Science and Network Security, VOL.10 No.7, July 2010.

[91] Bin Tang, Michael Shepherd, Malcolm Heywood and Evangelos Milios. Comparing and combining dimension reduction techniques for efficient text clustering. Faculty of Computer Science, Dalhousie University, Halifax, Canada, 2005.

[92] Elias Showk and Julian Bilcke. Mapping Wikileaks' Cablegate using Python, mongoDB, Neo4J and Gephi. FOSDEM 2011.

[93] Liu B (2007) Web data mining: exploring hyperlinks, contents and usage Data. Springer, Heidelberg.

[94] Brin S, Page L (1998) The anatomy of a large-scale hypertextual Web Search Engine. Computer Networks 30(1–7):107–117.

[95] Raffael Marty. Applied security visualization. Addison Wesley 2009
[96] Golub GH, Van Loan CF (1983) Matrix computations. The Johns Hopkins University Press
[97] Rada Mihalcea and Paul Tarau. TextRank: Bringing Order into Texts. Departament of computer science, Univesity of nort texas.
[98] Luis Torgo. Data Mining with R,CRC Press 2011.
[99] Jiawei Han and Micheline Kamber. Data Mining Concepts and Techniques, second edition, Elsevier 2006.
[100] Breunig, M., Kriegel, H., Ng, R., and Sander, J. (2000). LOF: identifying density-based local outliers. In ACM Int. Conf. on Management of Data,pages 93–104