Extracción de Información Semántica a Partir de Categorías de Texto Estado del Arte

Andrés Romero Rodríguez
Maestría en Ingeniería de Sistemas y Computación
Universidad Nacional de Colombia
Bogotá, Colombia
caromeroro@unal.edu.co

***Abstract*— Debido al aumento considerable en la cantidad de información que manejan tanto las empresas como las personas, se hace necesario contar con mecanismos que permitan organizar dicha información de forma que se pueda consultar posteriormente con facilidad y obtener rápidamente la información precisa que se esté buscando. Este artículo presenta una revisión de las técnicas utilizadas en el área de extracción de información enfocada a la obtención de contenido semántico a partir de categorías de texto que permita obtener conocimiento adicional al explicito en los documentos mediante descriptores o información asociada a las categorías. Esta información semántica está representada principalmente mediante palabras clave, redes semánticas y ontologías.**

***Index Terms*— Document Classification, Document Clustering, Information Retrieval, Information Extraction, Ontology Generation, Semantic Networks, Semantic Web.**

I. Introducción

Actualmente, el ritmo con el que se genera nueva información crece considerablemente a diario, de este modo, tanto las empresas como las personas, tienden a manejar volúmenes de datos enormes ya sea en forma de documentos de trabajo, información corporativa, correo electrónico, etc. Es por esto que se hace necesario el contar con herramientas que permitan manejar esta cantidad de información de forma confiable, segura y eficiente de modo que este gran volumen se convierta en una ayuda importante en la cotidianidad de las personas en lugar de ser un dolor de cabeza por su difícil manejo.

Las herramientas con las que se cuenta actualmente, no solucionan del todo el problema de la organización de la información, debido a que simplemente proveen un medio para su almacenamiento ya sea en forma de bases de datos, sistemas de gestión bibliográficos o manejadores de correo electrónico; pero aunque el acceso a dicha información se ha hecho muy eficiente, aún no se cuenta con técnicas adecuadas que permitan aprovechar el conocimiento tácito contenido en dicha información almacenada para mejorar el desempeño de los usuarios en las tareas cotidianas que involucran la utilización de esta información. Es por esto que el manejo de estos volúmenes de datos se ha vuelto un problema serio no solo a nivel corporativo sino a nivel personal, puesto que las personas cada día tienen que lidiar más con esta información que se hace inmanejable, desde el mismo uso del correo electrónico hasta el manejo de información en su empresa.

En los años recientes, se han desarrollado técnicas y metodologías encaminadas al manejo adecuado de esta información; entre ellos se encuentran los sistemas de clasificaición y categorización de texto, que buscan asignar cada uno de los documentos con que se cuenta en una de un conjunto de categorías predefinidas; técnicas de extracción de información que buscan obtener información relevante para el usuario en un determinado tópico a partir de un conjunto de documentos genérico; técnicas de semantic web, enfocadas a la organización de dicha información de modo que se conserve el conocimiento contenido en ella.

El presente artículo explora el campo de la obtención de información semántica a partir de categorías de texto, es decir, a partir de un conjunto de documentos agrupados en categorías que los relacionan en un contexto específico, se requiere obtener conocimiento de cada una de estas categorías tal como una descripción, un conjunto de palabras clave, etc. que permita acceder posteriormente de forma fácil a la información y se tenga un valor agregado que permita entender un poco mejor, por parte del usuario, el contenido de los documentos.

En la sección II se aborda el proceso de administración del conocimiento y se presentan las características semánticas principales que se deben tener en cuenta en este proceso; posteriormente en la sección III se hace una revisión de las técnicas utilizadas para la obtención automática de dicha información semántica. En la seción IV se visualiza el futuro de las técnicas de obtención y representación del conocimiento de modo que sea accesible para los usuarios, finalmente, en la sección V se dan algunas conclusiones del estudio.

II. Información Semántica

II-A. Representación del Conocimiento

La información semántica se refiere al conocimiento tácito que está contenido en los documentos, dicho conocimiento generalmente está oculto en el documento y no es fácilmente accesible; se deben buscar técnicas que permitan encontrar y representar el conocimiento almacenado en los documentos, en [11] se definen 5 principios que debe cumplir una representación del conocimiento, estos son:

1. *Una representación del conocimiento es un sustituto*: esto significa que al interior de la entidad que almacena

el conocimiento se presenta un proceso de razonamiento en el cual, este conocimiento almacenado es un sustituto de los objetos reales del mundo.
2. *Una representación del conocimiento es un conjunto de acuerdos ontológicos*: todas las representaciones que se puedan tener del conocimiento poseen cierto grado de error, el cual puede llevar a efectos no deseados en el proceso de manipulación del conocimiento. Es por esto que se debe llegar a un acuerdo acerca de las representaciones de modo que se trate de evitar ese error.
3. *Una representación del conocimiento es una teoría fragmentada de razonamiento inteligente*: esto se basa en la idea de que el conocimiento es la base para el razonamiento inteligente en los humanos. Al tener una representación de dicho conocimiento, esta no será del todo completa debido a las limitaciones inherentes de los sistemas de cómputo; por esto no se puede esperar un razonamiento completamente inteligente, sino que se debe fragmentar.
4. *Una representación del conocimiento es un medio para computación eficiente*: estas representaciones típicamente ofrecen un conjunto de ideas de modo que se facilite el proceso de inferencia, la efectividad de este proceso depende completamente de la eficiencia con que pueda ser procesado el conocimiento almacenado.
5. *Una representación del conocimiento es un medio de expresión humana*: es el medio para expresarse acerca del mundo; el medio para expresar a las máquinas conceptos acerca del mundo.

Dadas las características que debe cumplir una representación del conocimiento, se debe obtener dicho conocimiento a partir de un conjunto de documentos que contengan información relacionada acerca de un dominio particular, para esto, se deben usar técnicas de agrupamiento y clasificación de los documentos, así como metodologías que permitan representar y procesar el conocimiento almacenado en tales documentos, esta información se representará típicamente mediante conjuntos de palabras clave, redes semánticas y ontologías; adicionalmente es útil contar con información previa en el contexto de interés posiblemente utilizando ontologías de dominio público como en [16].

II-B. Administración del Conocimiento

Una vez se obtiene dicho conocimiento, se deben encontrar técnicas que permitan administrarlo de forma que sea útil para aquellas personas que requieren utilizarlo, de modo que se obtenga un valor agregado y mejore la eficiencia de los procesos involucrados y que requieran acceso a tal conocimiento. En [8] se presentan metodologías encaminadas a la administración del conocimiento al interior de las organizaciones, de modo que dicho conocimiento sea fácilmente accesible a las personas al interior de la organización, una ventaja importante de este enfoque es que toma en cuenta aspectos sociales. Las fases propuestas para llegar a un manejo apropiado del conocimiento son las siguientes:

1. *Obtención del conocimiento*: se refiere a la recolección y selección del conocimiento que requiere ser manejado.
2. *Organización y estructuración*: se debe imponer una estructura al conocimiento adquirido de modo que se pueda manejar eficientemente
3. *Refinamiento*: corregir, actualizar, adicionar o eliminar componentes de dicho conocimiento, en otras palabras, mantener el conocimiento apropiadamente.
4. *Distribución del conocimiento*: hacer disponible este conocimiento a los usuarios que lo requerian.

El objetivo final de este proceso de administración del conocimiento es llegar a la obtención de soluciones inteligentes a problemas difíciles a los que se enfrentan los usuarios y para los cuales este conocimiento es de utilidad si es aprovechado adecuadamente por los usuarios; este aprovechamiento se logrará gracias a una buena administración.

A partir del conocimiento obtenido se puede dar respuesta a algunas preguntas que son de interés para el usuario de dicha información, por ejemplo en [19] se presenta una técnica para dar explicaciones en el contexto de la clasificación de documentos, los resultados a los que se quiere llegar son:

- por qué se generan determinados clusters?
- como se relacionan varios clusters entre sí?
- explicaciones de cada cluster mediante jerarquías semánticas.

Además, en este proceso se han utilizado técnicas denominadas *Conceptual Clustering* para proporcionar descripciones de cada cluster, para esto se extraen de cada cluster las características mas relevantes encontradas en los documentos que hacen parte de dicha categoría: en [20] se presenta una técnica de dos etapas para el agrupamiento de documentos y además para obtener una descripción de cada uno de los clusters obtenidos. Las fases del algoritmo son las siguientes:

1. Clustering Tradicional: Utilizando una variante de K-means llamada Bisecting K-means se obtiene un conjunto de clusters que se pasarán a la siguiente etapa.
2. Clustering Conceptual: Se utilizan los clusters obtenidos en la fase anterior junto con un conjunto de tesauros para obtener descripciones de los clusters usando el algoritmo *Formal Concept Analysis*

En [10] se establecen los dos elementos principales en la adquisición de un lenguaje que sea de utilidad en la administración del conocimiento que se espera almacenar, estos son:

1. Construcción de un léxico, que es el conocimiento acerca de las palabras.
2. Generar relaciones entre las palabras del léxico mediante una red semántica o una ontología.

El conocimiento extraido de los documentos debe ser representado adecuadamente, de modo que se cumplan las características descritas anteriormente; este artículo se concentra en la extracción y representación de dicho conocimiento en forma de palabras clave, redes semánticas y ontologías como se muestra en la figura 1.

II-C. Redes Semánticas

Las redes semánticas surgen como un medio para almacenar conocimiento semántico, en un enfoque inicial, este

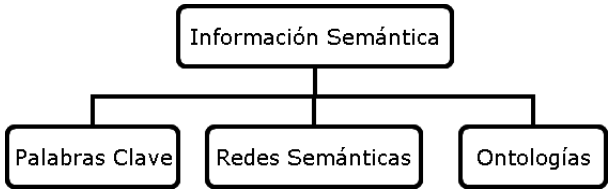

Fig. 1. Tipos de Representación del Conocimiento

conocimiento se basaba en jerarquías de herencia, pero posteriormente, se extendió el concepto de modo que se cubriera cualquier tipo de representación gráfica [17]. Estas redes se basan en la conexión de nodos que están relacionados mediante alguna característica semántica. Los componentes de una red, se pueden dividr en 3 categorías como en [3]:

- Objetos
- Eventos
- Relaciones

En donde los objetos representan todas las entidades tanto físicas como abstractas (ideas, números, etc), para cada objeto, se define su propiedad principal como el tipo al cual pertenece. Los eventos representan acciones, además se definen predicados que modifican la descripción de un evento adicionando un rol. Finalmente, las relaciones conectan objetos de diversos tipos, así como eventos que involucran dichos objetos.

Dichas redes son de gran utilidad puesto que proporcionan información acerca de los conceptos y las relaciones que existen entre ellos, estas relaciones pueden ser de diversos tipos, por ejemplo contenencia, herencia, etc. la característica común es que siempre representan una conexión semántica entre diversos conceptos que son de interés.

II-D. Ontologías

Una ontología se define como una *especificación formal y explícita de una conceptualización compartida* [8]; lo cual significa que esta representa un entendimiento común de algún dominio de modo que pueda ser comunicado tanto entre personas, como entre computadores. Las características que una ontología cumple, son las siguientes:

- Es un modelo abstracto de algún fenómeno del mundo, en el cual se identifican los conceptos relevantes.
- El tipo de conceptos usado, así como las restricciones que aplican sobre ellos, están definidos explícitamente.
- La ontología debe poder ser interpretada por un computador.
- La ontología captura conocimiento consensual, es decir, el resultado de un esfuerzo colaborativo de expertos en el dominio.

Las ontologías pueden ser expresadas utilizando diversos niveles de formalidad, sin embargo, existen cuatro categorías que son las formas más comunes de expresarlas [12]:

1. Altamente informal: escritas usando lenguaje natural
2. Semi-informal: Restringidas y estructuradas utilizando lenguaje natural
3. Semi-formal: Utilizando un lenguaje definido formalmente
4. Rigurosamente formal: semi-formal, incluyendo además teoremas y pruebas

Se debe encontrar el nivel de formalidad adecuado al momento de crear una ontología, puesto que el conocimiento representado en ella debe ser útil para los usuarios de la ontología y, además, esta debe ser usada para soportar el diseño y desarrollo de aplicaciones que hagan uso de este conocimiento, la mayoría de los cuales son sistemas basados en agentes para extracción de información en la web; dichos agentes deben ser capaces de compartir una ontología común para el acceso a la información.

Las ontologías puede ser clasificadas de acuerdo al nivel de formalidad en el que están escritas [12]:

1. Catálogo: una lista simple de términos, sin definiciones ni axiomas
2. Catálogo con definiciones: un catálogo con definiciones en lenguaje natural
3. Taxonomía: una colección de conceptos con un orden parcial
4. Taxonomía Axiomatizada: una taxonomía con axiomas
5. Biblioteca contextual: un conjunto de taxonomías axiomatizadas con relaciones entre ellas.

En [21] Se presenta DOGMA, un framework para el manejo de ontologías, a partir de este framework, se consideran aspectos relacionados con la escalabilidad y reusabilidad de las ontologías, dado que estas se pueden extender y unificar fácilmente a partir de ontologías existentes previamente; así mismo, se presentan algunos ejemplos relativos al uso e integración de diversas ontologías relacionadas, tomando como base un contexto de alquiler de vehículos.

II-E. Análisis Semántico Latente

Cuando se representan los documentos a partir de las palabras que estos contienen, se está dejando de lado información adicional contenida en la estructura del documento, adicionalmente, se presentan fenomenos como la sinonímia (palabras diferentes son usadas para describir el mismo concepto) y la polisemia (una palabra puede tener diferentes significados de acuerdo al contexto), los cuales tienen un impacto importante en la efectividad de los métodos de extracción de información cuando los documentos son representados únicamente mediante un vector de palabras.

Existen dos tipos de enfoques para representar el espacio de características de los documentos, el primero es llamado *selección de características*, que consiste en seleccionar un subconjunto de las características más representativas del espacio original. el otro enfoque, llamado *extracción de características*, consiste en transformar el espacio original en un espacio de características más pequeño para reducir la dimensión; la ventaja de este segundo enfoque es que además de reducir la dimensionalidad del conjunto de datos, también soluciona de forma apropiada los problemas de sinonimia y polisemia [24]. La técnica mas representativa de extracción de características es el análisis semántico latente.

El análisis semántico latente es una técnica estadística que intenta estimar la estructura oculta inherente a los documentos dada por los términos contenidos relacionados a conceptos.

Para llevar a cabo este proceso, se emplea una técnica denominada SVD (*Singular Value Decomposition*), que está basada en el álgebra lineal, con el objetivo de descubrir los patrones asociativos mas importantes entre las palabras y los conceptos [5]; esto lo hace mapeando tanto los documentos como los términos en una representación en el llamado *espacio semántico latente*; el objetivo de llevar los documentos y términos a este espacio, es que en él las similitudes entre documentos o entre documentos y consultas se puede estimar de forma más apropiada que en el espacio original. Otro punto importante, es que los documentos que comparten palabras relacionadas tendrán una representación similar en el espacio semántico latente, aún si estos no tienen palabras en común [18].

En LSA, se representan los documentos como una matriz X de la siguiente forma:

$$X = [x_1, x_2, ..., x_n]^T$$

en donde cada vector x_i contiene las ocurrencias de los términos en el documento i, la técnica consiste en calcular el SVD de X y posteriormente asignar a cero todos los valores excepto los K mayores; los documentos son entonces mapeados en el espacio semántico seleccionando las direcciones de máxima covarianza [5].

II-F. Palabras Clave

Las palabras clave brindan información resumida del contenido de un documento, a medida que la cantidad de documentos disponibles aumenta, esta información se hace cada vez más importante puesto que pueden ser usadas en sistemas de extracción de información como punto de partida, en lugar de utilizar todo el contenido del documento; como base para sistemas de indexamiento; o como medio para realizar búsquedas dentro de grandes colecciones de documentos. Las palabras clave son útiles dado que son independientes entre ellas y pueden ser interpretadas individualmente [30].

La representación tradicional de los documentos se hace utilizando el llamado *bag of word*, que consiste en crear un vector para cada documento, en donde cada posición de dicho vector corresponde a la representación de una palabra.

$$A = (a_{ij})$$

En donde cada fila i de la matriz A representa un documento, y la posición (i, j) representa el peso de la palabra j en el documento i. Las diferentes representaciones que utilizan este enfoque se diferencian en la forma en que se calcula el peso de las palabras dentro del documento, las representaciones mas usadas son [1]:

- *Boolean weighting*: el peso es 1 si la palabra pertenece al documento o 0 en caso contrario
- *Word frequency weighting*: el peso es la frecuencia de la palabra en el documento
- *tf*idf weighting*: asigna el peso de la palabra en proporción al número de ocurrencias de la palabra en el documento y en proporción inversa al número de documentos en la coleción en los cuales aparece la palabra.

Estas representaciones tradicionales de los documentos como un vector de palabras con un peso asociado no siempre son apropiadas puesto que está limitada a la aparición explicita de los términos en el conjunto de documentos, además, no se aprovecha el conocimiento acerca del mundo que tienen los usuarios [16]; se deben explorar entonces alternativas que aprovechen este conocimiento de modo que no se dependa únicamente del contenido sintáctico de los documentos. Para tratar este problema, se hace uso de palabras clave que están asociadas al dominio en el cual se trabaja y que proporcionan información adicional que posiblemente no esté contenida en los documentos.

III. Obtención de Información Semántica

En el proceso de obtención de información semántica, es importante tener en cuenta los aspectos mencionados en la sección anterior, además, se debe partir de una representación y administración de los documentos que permita manipularlos adecuadamente para lograr este objetivo.

En [25] se hace un estudio acerca de las características lungüísticas complejas que se deben tener en cuenta para mejorar el desempeño de los clasificadores de texto. Generalmente, las representaciones utilizadas no tienen en cuenta dichas características; por lo que se plantean nuevas formas de representación, además, se hace una experimentación utilizando estas nuevas representaciones.

Otras formas de representación del conocimiento son exploradas en [29]. Aquí se hace una revisión de un enfoque para la representación de conocimiento llamado FCA (Formal Concept Analysis), este enfoque tiene sus orígenes en las matemáticas como la formalización del concepto "Concepto", se muestra además la relación entre este concepto matemático y su aplicación en las ciencias de la computación. FCA se aborda como una técnica de representación del conocimiento. Se utilizan técnicas de extracción y procesamiento de conocimiento conceptual, las cuales se dividen en:

- Adquisición
- Representación
- Inferencia
- Comunicación

III-A. Construcción de Redes Semánticas

Como se definió en la sección anterior, las redes semánticas son un medio para relacionar conceptos entre si relativos al contexto de interés; para obtener esta información, es necesario tener en cuenta la representación de los documentos y la forma en que esta representación permite extraer dichas relaciones. En [28] se examinan formas alternativas para la representación de documentos en el contexto de la clasificación de texto, mas allá del tradicional bag of words (en donde cada palabra corresponde a una característica); dicha representación nueva se basa en relaciones sintácticas y semánticas entre las palabras (frases, sinónimos, etc).

En [17] se presenta un enfoque orientado a la representación semántica para modelar las probabilidades con que las palabras aparecen en los diferentes contextos y a partir de allí, se busca capturar las relaciones entre palabras. En primera instancia, se obtiene una representación semántica en un espacio de dimensión menor al original, a partir de esta representación, es

posible obtener redes semánticas básicas, o redes semánticas bipartitas, las cuales constan de nodos de 2 tipos, un tipo representa tópicos y el otro representa palabras relacionadas con dichos tópicos, una característica importante de este tipo de redes es que solo pueden existir conexiones entre nodos de distinto tipo, de este modo se relacionan palabras con conceptos.

En [6] se presenta un algoritmo para la construcción de redes semánticas a partir de documentos de texto, la idea es llegar a una representación de cada documento como una red semántica que contenga los conceptos más importantes relacionados en el documento. Para la construcción de esta red, se utiliza WordNet como base de conocimiento, la cual es vista como una red semántica general de la cual se van a obtener conceptos que serán útiles para la construcción de la red que representará el documento. El algoritmo consiste en detectar los conceptos que pueden ser palabras o frases completas a partir de WordNet y que se encuentren en el documento, posteriormente se calculan las frecuencias de dichos conceptos en el documento y se le asigna un peso a cada uno; finalmente, se comparan los conceptos encontrados entre sí para obtener una red que relacione aquellos conceptos que son similares.

En [14] se presenta un modelo para la generación auntomática de árboles de clasificación que corresponden a un modelo jerárquico de lenguaje. Se muestra que es importante tener en cuenta que la frecuencia de las palabras en un documento está fuertemnte relacionada con el contexto del documento. El enfoque general consta de 2 fases:

1. Agrupamiento de los documentos: se utiliza una técnica de clustering no supervisado para agrupar documentos que tienen tópicos relacionados
2. Generación del árbol: en esta etapa se utiliza un algoritmo de clustering jerárquico aglomerativo, el árbol queda construido de modo que los documentos similares van a quedar juntos.

III-B. Construcción de Ontologías

A partir de información textual o categorizada, se puede extraer información en forma de ontologías que representen el conocimiento almacenado en los documentos pertenecientes a cada una de las categorías. En [9] se presenta un enfoque para el acceso a la información contenida en documentos de texto a partir de los tópicos encontrados en dichos documentos. La idea básica, es combinar técnicas de segmentación de tópicos y técnicas de extracción de información. La extracción se realiza a partir de una ontología construida previamente, a partir de la cual se utilizan técnicas de extracción de información para conectar segmentos de texto obtenidos con los conceptos dados por la ontología. Finalmente lo que se obtiene es un conjunto de subdocumentos mas pequeños para cada documento que tienen tópicos homogeneos.

El proceso de construcción de una ontología requiere un análisis del dominio, el cual se implementa de la siguiente forma [27]:

1. Identificar cuidadosamente el vocabulario usado para describir los conceptos relevantes en el dominio
2. Codificar definiciones completas y rigurosas acerca de los términos (conceptos) en el vocabulario
3. Caracterizar las relaciones conceptuales entre dichos términos

En [27] se definen además las tres características necesarias para construir ontologías útiles:

- **Cubrimiento:** La ontología debe estar suficientemente poblada (para los propósitos de la aplicación). Se necesitan herramientas para soportar las tareas de identificación de los conceptos relevantes y las relaciones entre ellos
- **Consenso:** Debe existir un acuerdo en los aspectos básicos del dominio, este consenso se logra entre aquellas personas involucradas en el manejo de la ontología
- **Accesibilidad:** Se requieren herramientas que permitan una integración fácil de la ontología dentro de la aplicación.

Adicionalmente, se presenta una herramienta (*OntoLearn*) para la extracción de conocimiento a partir de documentos electrónicos y se describe una metodología para la construcción automática de ontologías utilizando dicho conocimiento. Los pasos básicos de esta metodología son:

1. Extracción de terminología candidata a partir de documentos pertenecientes al dominio
2. Filtrado de terminología del dominio, para esto se utilizan documentos relativos a dominios que contrasten con el de interés
3. Interpretación semántica de los términos obtenidos
4. Identificación de relaciones taxonómicas y de similaridad
5. Generación de la ontología

En [13] se presenta una metodología para construir ontologías a partir de componentes de cada categoría visualizados mediante un mapa auto-organizativo (SOM); para esto se obtienen las palabras que aportan mayor información a cada categoría y con ellas se consruye el SOM del que se obtiene una ontología para dicha clase. El esquema utilizado efectúa los siguientes pasos:

1. Se obtiene un conjunto de documentos relacionados
2. Se realiza un pre-procesamiento de dichos documentos eliminando aquellas palabras que aportan poca información
3. Se obtienen las palabras base y se asignan pesos a cada uno de los términos usando tf*idf y se representa el documento como un vector
4. Con los vectores individuales de cada documento, se construye el *espacio del documento*
5. Se construye una red SOM a partir del espacio de documentos generado
6. Se crea la ontología a partir de los componentes visualizados en el mapa

En [12] se muestra una metodología para construir ontologías automáticamente a partir de la información contenida en los documentos que están siendo evaluados y en documentos adicionales. El enfoque que se le da a la construcción de dicha ontología es dirigido por los intereses de cada usuario, así, cada uno podrá tener una ontología diferente de acuerdo a lo que esté buscando. El esquema que se sigue es la utilización

de un conjunto de palabras clave generados por el usuario junto con la información contenida en el documento para verificar si esta clasificación representa una abstracción mas general que el propio contenido del documento. La ontología se construye en principio realizando un conteo de las palabras que generalizan las palabras clave dadas por el usuario y sus sinónimos. Posteriormente se aplica PCA para encontrar las palabras mas importantes. Aquí, se definen además, las características básicas que se deben tener en cuenta al momento de construir una ontología, estas son:

1. Claridad: la ontología debe proporcionar el significado esperado de los términos definidos
2. Coherencia: los axiomas deben ser consistentes lógicamente
3. Exensibilidad: cuando se definen nuevos términos, no debe ser necesario revisar las definiciones existentes
4. Suficiencia: la ontología debe ser suficiente para soportar el propósito para el cual es construida
5. Principio de distinción: las clases deben ser disyuntas
6. Diversificación: incrementa el poder proporcionado por múltiples mecanismos de herencia
7. Minimización de distancias semánticas: las clases similares deben estar agrupadas juntas
8. Codificación: la ontología debe estar especificada en un nivel de conocimiento dado sin depender de una codificación particular de símbolos

III-C. Análisis Semántico Latente

El análisis semántico latente se basa en el algoritmo de descomposición de valores singulares (SVD), que genera una representación como una matriz de relaciones entre términos. Una vez se han obtenido estos valores, es necesario contar con mecanismos que permitan entenderlos apropiadamente de modo que se puedan utilizar estas relaciones como ayuda en el proceso de extracción de información semántica. En [26] se presenta un framework para detectar las correlaciones entre las palabras a partir de los valores generados en la matriz de términos mediante SVD, adicionalmente, se presenta un enfoque para la detección de patrones en los datos; dichas correlaciones entre términos proporcionan un modelo para alcanzar un entendimiento semántico del LSA.

En [4] se presenta una variación al algoritmo SVD (Singular Value Decomposition) utilizado en LSA (Latent Semantic Analysis), el algoritmo propuesto difiere del SVD en cuanto a que éste tiene en cuenta los documentos que son anormales, los cuales son clasificados como ruido por SVD y por lo tanto no son tenidos en cuenta, este algoritmo los tiene en cuenta aplicando un escalamiento a dichos datos de modo que al momento de hacer la descomposición tengan un efecto significativo sobre los demás datos. Esto se hace con el fin de mejorar la medida de similitud entre documentos. Se plantea además que se puede llegar a encontrar esos factores de escalamiento de los vectores anormales de manera dinámica, lo cual mejoraría el desempeño del algoritmo.

En [18] se presenta una extensión al indexamineto semántico (LSI) puesto que este tiene deficiencias en cuanto a sus fundamentos estadísticos, la idea es mapear tanto los documentos como los términos a una representación en el espacio semántico, reduciendo así la dimensionalidad y facilitando la extracción de información.

En [24] se presenta un enfoque para el indexamiento semántico basado en relevancia, teniendo en cuenta una discriminación entre categorías, este método mejora el rendimiento del análisis semántico global puesto que este solo se concentra en la representación global de los documentos y no tiene en cuenta las diferencias entre las clases de los documentos.

Una forma intuitiva de describir la información contenida en los documentos es el uso de jerarquías, estas han sido utilizadas ampliamente como un mecanismo de clasificación, dado que es fácil entender la clasificación que generan y son útiles para resumir y organizar conjuntos de datos grandes [23].

Las jerarquías pueden ser creadas de modo que sean orientadas al documento u orientadas a los términos. Una jerarquía orientada al documento es aquella en la que los documentos son divididos en cada uno de los niveles de la jerarquía. Una jerarquía orientada a los términos utiliza los conceptos contenidos en los documentos para formar una estructura jerárquica. Existen dos técnicas principales en la generación automática de jerarquías basadas en términos:

- **Jerarquías de contenencia:** Se basa en la noción de contenencia de los términos. Dado un conjunto de documentos, algunos términos aparecerán en la mayoría de documentos, mientras que otros aparecerán solo en unos pocos. Algunos de los términos que aparecen más frecuentemente proporcionan gran información acerca del tópico general de los documentos; existirán algunos términos que definan ampliamente los tópicos, y otros que aparecen junto con un término general que explican aspectos referentes al tópico en cuestión. Las jerarquías de contenencia tratan de sacar provecho de este tipo de palabras. Este tipo de jerarquías tiene las siguientes características:
  - Posee un medio para asociar términos de modo que se reflejen los tópicos cubiertos en los documentos
  - Dentro de la asociación, un término padre es más general que un hijo
  - Un término contiene a todos sus descendientes de modo que se mantiene una relación transitiva
  - Un hijo puede tener más de un padre
- **Jerarquías Léxicas:** Otro enfoque para crear jerarquías es utilizando la estructura jerárquica de las frases que aparecen frecuentemente. Se deben seleccionar las frases que servirán como candidatas en la jerarquía léxica, posteriormente, estas se dividen en grupos basados en los términos que aparecen en las frases y se debe calcular la dispersión de cada uno de los términos. También es necesario examinar el número de documentos que involucran frases que contienen un determinado término.

III-D. Extracción de Palabras Clave

Las palabras y frases clave usualmente son seleccionadas manualmente. En muchos contextos académicos los autores son quienes asignan estas palabras clave a los documentos

que escriben. Este enfoque funciona adecuadamente mientras todos los documentos que se consideran tengan asignado un conjunto de palabras clave; pero en la práctica, no se alcanza esta situación dado que la cantidad de información es enorme y generalmente no está bien estructurada. Es por esto que se debe extraer dichas palabras clave automáticamente a partir del contenido del documento; para lograr esto existen dos enfoques principales [30]:

- **Asignación:** Se seleccionan las frases que mejor describan al documento de un vocabulario controlado. En su fase de entrenamiento, se asocia un conjunto de documentos con cada frase del vocabulario y se construye un clasificador por cada frase. Cada nuevo documento es procesado por cada clasificador y se asignan las frases adecuadas. Las únicas frases que pueden ser asignadas son aquellas que han sido seleccionadas en la fase de entrenamiento.
- **Extracción:** No se utiliza un vocabulario controlado, en lugar de ello las frases son seleccionadas automáticamente del texto mismo. Se emplean técnicas de extracción de información y procesamiento léxico para extraer frases que tengan una alta probabilidad de caracterizar el documento. En esta técnica los datos de entrenamiento únicamente son utilizados para ajustar los parámetros del algoritmo de extracción.

En [16] se presenta una técnica para la generación de características adicionales a las contenidas en los documentos, estas características constituirán un conjunto de palabras clave que se utilizarán para el procesamiento de los documentos junto con las palabras contenidas en los mismos. Esta generación de características se basa en un conocimiento específico del dominio y de sentido común, el cual está representado como ontologías que contienen cientos de conceptos. El generador de características analiza los documentos y los mapea en conceptos pertenecientes a las ontologías, lo cual induce un conjunto de caracterśticas adicionales a las contenidas en el documento.

En [7] se estudia una técnica de extracción de características conocida como *Word Clusters*, la cual busca generar grupos de palabras relacionadas en cuanto a las categorías que representan, para esto se utiliza el algoritmo *Information Bottleneck*, el cual genera representaciones compactas que permiten mejorar el procesamiento de los documentos. Estos clusters generados representan las características principales del documento y adicionalmente muestran una relación implicita entre los términos.

En [15] se presenta una técnica para la categorización de documentos y generación simultanea de keywords para cada categoría. Se presenta un algoritmo no supervisado basado en K-means en donde cada cluster está representado como un grupo de keywords. Se plantean 2 principales ventajas que tiene este enfoque:

- Los clusters generados van a tener un significado semántico mas apropiado
- Es posible generar, a partir de dichos keywords, una descripción de cada cluster automáticamente

En este algoritmo, cada cluster tiene un grupo de características y un peso asociado, a partir de los cuales se obtendrán posteriormente los keywords de cada categoría. En cada iteración del algoritmo se van ajustando los pesos lo que hace que al final se obtengan aquellos que son mas relevantes.

En [22] se introduce una técnica para la generación semiautomática de diccionarios temáticos mediante algoritmos de clasificación de texto; para estos algoritmos, se emplean técnicas de extracción de información. Además, se tiene en cuenta las asociaciones entre los términos y los temas, dichos términos son representados por un vector, además cada término tiene un tema asociado, en lugar de asociar este a un documento completo.

En [2] se presenta una metodología para adicionar componentes de procesamiento de lenguaje natural en tareas de clasificación de texto. Dado que las técnicas utilizadas tradicionalmente para llevar a cabo el procesamiento del texto con miras a la clasificación (Extracción de términos, asignación de pesos y reducción de la dimensionalidad), se llevan a cabo de una forma muy simple, es necesario contar con mecanismos que permitan adicionar este tipo de técnicas de lenguaje natural para mejorar el desempeño de los métodos de categorización. Para tratar este problema, el autor propone una metodología que adiciona tantas palabras como sea posible usar en la fase de clasificación, incluyendo términos que tienen baja frecuencia. Se introduce una extensi´on a la noción de td-idf que tiene en cuenta estas palabras poco frecuentes. Además de tener en cuenta palabras simples, se tienen en cuenta términos compuestos por varias palabras.

IV. Perspectivas

Como se ha mostrado en las secciones anteriores, el problema de extracción y representación del conocimiento ha sido abordado desde diversos enfoques, principalmente en cuanto a la generación de palabras clave, construcción automática de ontologías y, en menor medida, a la construción de redes semánticas. Estos enfoques brindan formas apropiadas de representar la información contenida en los documentos, así mismo facilitan el posterior procesamiento de este conocimiento por parte de un computador, lo cual ayudará a que esté disponible para los usuarios en la forma en que estos lo requieran.

Surgen entonces problemas en los cuales se debe profundizar el trabajo de modo que este proceso de administración del conocimiento sea efectivo en sus metas. las áreas en las que se ubican estos problemas son las mismas discutidas en las secciones anteriores:

- Extracción de características. Se ha mostrado que el tradicional *bag of words* no es apropiado para representar la información de los documentos puesto que no se tiene en cuenta las relaciones tácitas entre las palabras. Aquí se mostraron algunos enfoques alternativos para solucionar este problema, posiblemente el análisis semántico latente es la técnica que mas se aproxima a lo que debería ser una representación adecuada del conocimiento; pero aún se necesita un poder de expresividad mayor, de modo que se puedan conservar las estructuras semánticas contenidas

en el texto; esto se logra involucrando técnicas de procesamiento de lenguaje natural que darán mayores alcances a las representaciones, pero así mismo requerirán de metodologías de procesamiento más sofisticadas para manejar dicha información.

- Obtención del conocimiento: Aunque es claro que las técnicas de representación tales como redes semánticas y ontologías son apropiadas para almacenar la información contenida en los documentos, aun se debe trabajar en la generación automática de estas representaciones. En cuanto a las redes semánticas, el trabajo aunque es escaso ha producido resultados interesantes para dominios particulares, se debe trabajar entonces en mecanismos de generación automática de estas redes que sean independientes del contexto para llegar finalmente a que sean capaces de determinar el dominio de los documentos y alcanzar esa meta de dar explicaciones y descripciones de la información representada en los textos.
  En el área de la generación automática de ontologías existen dos enfoques principales:

  - Generación de ontologías
  - Generación de instancias de una ontología dada

  Los resultados que se han obtenido hasta el momento en estos dos enfoques aun están lejos de ser comparados con los que se pueden llegar a generar manualmente; se han explorado diversas técnicas para solucionar este problema, pero ninguna ha tenido un resultado suficientemente bueno para ser considerada una solución definitiva. Es necesario considerar otras técnicas de aprendizaje de máquina y procesamiento de lenguaje natural que ayuden en este objetivo.
- Adicionalmente, estas representaciones deben ser lo suficientemente flexibles para proporcionar la información que el usuario necesita en la forma en que la necesita, generalmente esta búsqueda de información está basada en una consulta que es comparada con los documentos para determinar cuales son apropiados; pero en un caso más general, estas consultas pueden ser mucho más elaboradas y llegar a convertirse en simplemente preguntas que el usuario hace, las cuales deben ser interpretadas adecuadamente y se debe encontrar la información relevante; en este caso no es posible tratar la consulta como un conjunto de palabras clave que se buscarán en los documentos, sino que se debe considerar una tarea de extracción de información tanto para la consulta, es decir, saber qué está buscando el usuario y para los documentos, saber en donde se encuentra aquello que se esté buscando.
- Finalmente, es necesario brindar a los usuarios información relevante de todo este proceso, se han presentado desarrollos enfocados a la explicación en el contexto de clasificación de texto, es decir, se trata responder a las preguntas de por qué se generan ciertos grupos o por qué un documento pertenece a un grupo. Estas explicaciones deben tender a dar información mas fácilmente legible por el usuario, por ejemplo descripciones de un grupo de documentos, o extracción de información particular a partir de múltiples fuentes.

V. Conclusiones

Se ha hecho una revisión de las características que se deben tener en cuenta para un proceso de administración del conocimiento, de modo que sea accesible a los usuarios y se han revisado los principales enfoques relativos a la obtención de información semantica a partir de categorías de texto. Estos enfoques se dividen en tres grandes grupos: Palabras clave, Redes semánticas y Ontologías. Se estudiaron las características que tiene cada uno de estos tipos de representación del conocimiento y se revisaron los principales enfoques orientados a la construcción automática de los mismos.

El resultado de todo este proceso de representación y administraciíon del conocimiento es hacerlo accesible a todos los usuarios que lo requieran, ya sea a nivel corporativo o personal. En las origanizaciones es necesario puesto que constantemente se están generando documentos de trabajo, manuales, o en general cualquer tipo de información corporativa que requiere ser almacenada y procesada de una forma eficiente y, finalmente, darla a conocer a las personas tanto internas como externas de la organización en niveles de detalle que dependen de los intereses o la importancia del usuario. A nivel personal es útil puesto que actualmente se genera información en muchas fuentes, una de las mas importantes es el correo electrónico, el cual no debe convertirse en un dolor de cabeza por la cantidad de información sino que debe convertirse en un medio que facilite el trabajo diario de las personas.

Referencias

[1] K. Aas and L. Eikvil. Text categorisation: A survey., 1999.

[2] A.Ñ. Aizawa. Linguistic techniques to improve the performance of automatic text categorization. In *NLPRS*, pages 307–314, 2001.

[3] J. F. Allen and A. M. Frisch. What's in a semantic network? In *ACL Proceedings, 20th Annual Meeting*, pages 19–27, 1982.

[4] R. K. Ando. Latent semantic-space: iterative scaling improves precision of inter-document similarity measurement. In *SIGIR*, pages 216–223, 2000.

[5] P. Baldi, P. Frasconi, and P. Smyth. *Modeling the Internet and the Web: Probabilistic Method and Algorithms*. John Wiley, 2003.

[6] M. Baziz, M. Boughanem, and N. Aussenac-Gilles. Semantic networks for a conceptual indexing of documents in IR. In *The Seventh International Symposium On Programming and Systems (ISPS) , Algiers*, pages 213–224. ,, 9-11 mai 2005.

[7] R. Bekkerman, R. El-Yaniv, N. Tishby, and Y. Winter. Distributional word clusters vs. words for text categorization. *Journal of Machine Learning Research*, 3:1183–1208, 2003.

[8] V. R. Benjamins, D. Fensel, and A. Gómez-Pérez. Knowledge management through ontologies. In *PAKM*, 1998.

[9] C. Caracciolo, W. R. van Hage, and M. de Rijke. Towards topic driven access to full text documents. In *ECDL*, pages 495–500, 2004.

[10] S. Chakrabarti. *Mining the Web. Discovering Knowledge from Hypertext Data*. Morgan Kaufmann Publishers, 2002.

[11] R. Davis, H. Shrobe, and P. Szolovits. What is a knowledge representation. *AI Magazine*, 14(1):17–33, 1993.

[12] D. Elliman. Automatic derivation of on-line document ontologies, July 26 2001.

[13] D. Elliman and J. R. G. Pulido. Visualizing ontology components through self-organizing maps. In *IV*, page 434, 2002.

[14] R. Florian and D. Yarowsky. Dynamic nonlocal language modeling via hierarchical topic-based adaptation. In *37th Annual Meeting of the Association for Computational Linguistics*, pages 167–174, 1999.

[15] H. Frigui and O.Ñasraoui. Simultaneous categorization of text documents and identification of cluster-dependent keywords, Apr. 07 2002.

[16] E. Gabrilovich and S. Markovitch. Feature generation for text categorization using world knowledge. In *IJCAI*, pages 1048–1053, 2005.

[17] T. L. Griths and M. Steyvers. A probabilistic approach to semantic representation, Apr. 29 2002.

[18] T. Hofmann. Probabilistic latent semantic indexing. pages 50–57.

[19] A. Hotho, S. Staab, and G. Stumme. Explaining text clustering results using semantic structures. In *PKDD*, pages 217–228, 2003.

[20] A. Hotho and G. Stumme. Conceptual clustering of text clusters, May 23 2002.

[21] M. Jarrar and R. Meersman. Scalability and knowledge reusability in ontology modeling. In V. Milutinovic, editor, *Proceedings of the International conference on Infrastructure for e-Business, e-Education, e-Science, and e-Medicine*, volume SSGRR2002s, Rome, Italy, 2002. SSGRR education center.

[22] A. Lavelli, B. Magnini, and F. Sebastiani. Building thematic lexical resources by term categorization. In *SIGIR*, pages 415–416, 2002.

[23] D. Lawrie and W. B. Croft. Discovering and comparing topic hierarchies, Oct. 13 2000.

[24] T. Liu, Z. Chen, B. Zhang, W.-Y. Ma, and G. Wu. Improving text classification using local latent semantic indexing. In *ICDM*, pages 162–169, 2004.

[25] A. Moschitti and R. Basili. Complex linguistic features for text classification: A comprehensive study. In *ECIR*, pages 181–196, 2004.

[26] W. M. Pottenger and P. D. Detecting patterns in the LSI term-term matrix. Technical report, Sept. 25 2002.

[27] A. Scime. *Web Mining: applications and techniques*. Idea Group, 2005.

[28] S. Scott and S. Matwin. Feature engineering for text classification. In *ICML*, pages 379–388, 1999.

[29] G. Stumme. Formal concept analysis on its way from mathematics to computer science. In *ICCS*, pages 2–19, 2002.

[30] I. H. Witten, G. W. Paynter, E. Frank, C. Gutwin, and C. G. Nevill-Manning. KEA: Practical automatic keyphrase extraction. *CoRR*, cs.DL/9902007, 1999.