Capítulo 1

Introducción

Se dice que estamos en la era de la información debido a que actualmente se tiene acceso a prácticamente toda la que se necesita. La facilidad para los usuarios de publicar sus documentos en la Web ha provocado que exista una cantidad agobiante de información y se hace difícil su administración y recuperación. Por lo anterior, debemos aprender a filtrarla y sintetizarla [6]. El área de recuperación de información ha tenido mucho auge porque ayuda a cumplir este propósito.

Como se menciona en [25], la recuperación de información es el proceso de manipular un texto y almacenarlo, para luego recuperarlo y con los resultados obtenidos se informa al usuario si el documento que contiene la información solicitada está disponible o no.

La recuperación de información se vale de varias técnicas para llevar a cabo la búsqueda de documentos relevantes a una consulta en algún repositorio, por ejemplo a una base de datos. Para que esta búsqueda tenga éxito, es muy importante que los documentos almacenados estén bien clasificados y agrupados. Dicha clasificación actualmente se lleva a cabo con métodos no automatizados que asignan palabras clave relevantes al tema en torno al que gira el texto. También existen métodos automatizados, sin embargo o son muy complejos porque se basan en técnicas matemáticas, probabilísticas o de vectorización, o bien, no son confiables porque en algunos casos

sólo toman en cuenta la repetición de palabras en el texto, cuando se sabe que en la mayoría de los casos lo que le da sentido a estas palabras es el contexto.

Los métodos mencionados anteriormente no aplican técnicas de interpretación de lenguaje natural y, por lo tanto, ninguna teoría lingüística, aunque existen algunos métodos que sí lo hacen, pero éstos se tornan muy complejos porque para su implementación se requieren estructuras de datos y algoritmos difíciles de diseñar y formalizar, así como de la intervención de expertos humanos para el preprocesamiento y postprocesamiento. Además la aplicación de los métodos que analizan las estructuras gramaticales se restringe por el hecho de que varían mucho con el idioma y el dominio. Por ejemplo, en los casos en los que existe ambigüedad en la función de una palabra dentro de la oración, pueden no ser muy eficaces. Se piensa que en general se obtendrían mejores resultados en la interpretación de documentos, y por lo tanto en su clasificación, si se aplican técnicas que procesen textos en lenguaje natural de manera similar a la interpretación del lenguaje que llevan a cabo los humanos.

Existen muchas teorías lingüísticas que buscan explicar el procesamiento del lenguaje en los humanos y generar un modelo matemático de éste. Algunos de estos enfoques se han tratado de aplicar de manera computacional, por ejemplo, dentro de las ideas de Noam Chomsky, la teoría de la sintaxis así como la teoría de la gramática categorial de Ajdukiewicz [17]. Basándose solamente en las reglas de estos modelos, se puede generar un sinnúmero de estructuras a partir de un vocabulario, aunque sólo un porcentaje pequeño de ellas se utiliza. Realmente pasa algo así con los hablantes no nativos de un idioma que construyen estructuras gramaticalmente correctas pero que no suenan bien porque un nativo de éste utilizaría una estructura diferente. De esto podemos entender que, para muchas aplicaciones, no es necesario procesar todas las posibles combinaciones si en realidad no son usadas, y este tipo de modelos resultaría innecesariamente complejo.

La teoría formulaica es una teoría lingüística reciente que se piensa que será utilizada para aplicaciones computacionales, por ejemplo para el análisis de textos no estructurados [21]. Se piensa que bajo la premisa de que el ser humano no realiza un análisis tan complejo del lenguaje, puesto que tiene fórmulas pre-guardadas, se obtendrían buenos resultados simulando el procesamiento formulaico de los humanos. También se piensa que la aplicación de la teoría formulaica podría darnos algunas ventajas sobre los métodos tradicionales de análisis de textos.

1.1. Descripción del problema

Los métodos tradicionales de análisis de textos están basados en técnicas matemáticas (estadísticas y probabilísticas) y requieren de un preprocesamiento de textos muy complejo, porque trabajan sólo sobre textos estructurados. Es decir, si partimos de un texto no estructurado, se tiene que indexar, definir categorías gramaticales de las palabras, construir vectores, árboles de categorización, etc. Además es necesario un postprocesamiento que podría ser algo poco práctico, porque en algunos casos requiere de intervención humana, como por ejemplo en el caso de redes neuronales, cuyas reglas de decisión resultantes no son interpretables. Así, ese postprocesamiento tendría que ser hecho por un experto en el dominio que se asegure de que los resultados son buenos, y a través de su juicio determinar si el sistema se está comportando bien y si ya está suficientemente entrenado.

1.2. Objetivos

1.2.1. Objetivo general

- Construir una aplicación computacional de la teoría formulaica que obtenga buenos resultados dentro de la clasificación de textos no estructurados.

1.2.2. Objetivos específicos

1. Modelar o diseñar un algoritmo que clasifique los textos que tenga en principio una fase de entrenamiento y una de pruebas.

2. Compilar un corpus de textos para entrenamiento y otro para pruebas.

3. Seleccionar los métodos computacionales adecuados para la implementación de la teoría formulaica y desarrollar una metodología que los utilice.

4. Realizar análisis de textos e identificar las fórmulas propias de cada dominio (entrenamiento).

5. Construir una base inicial de fórmulas que se pueda ir mejorando al analizar cada texto.

6. Desarrollar una metodología que permita que el tamaño de las fórmulas sea variable dependiendo de cada dominio.

7. Clasificar textos de prueba haciendo una comparación entre éstos y la lista de fórmulas obtenidas en la fase de entrenamiento.

8. Utilizar la relevancia probabilística de las fórmulas a cada domino para desambiguar casos dudosos de clasificación.

9. Buscar una implementación que sea flexible, es decir independiente del idioma y dominio.

10. Se busca la sencillez en la implementación de todos los algoritmos que se utilicen.

1.3. Alcances

- La ejecución del algoritmo se realiza en dos fases; la primera es alimentada de los textos de entrenamiento, en base a los cuales construye un conjunto inicial de fórmulas. En la segunda fase el algoritmo tiene como entrada textos de prueba, los cuales debe clasificar con base en el conjunto de fórmulas obtenido en la fase anterior.

- Se revisaron las ventajas y desventajas de varios algoritmos para determinar cuál es el más adecuado en la identificación de fórmulas. Se pensó en algunas herramientas como SUBDUE [9] y un algoritmo ad-hoc, este último propuesto en este trabajo. El funcionamiento de ambos se explicará más adelante.

- Cada fórmula tiene un peso asociado a la relación que existe entre ésta y cierto dominio y así tener un criterio probabilístico acerca de la relevancia de una fórmula a un dominio.

- Se revisaron algunos algoritmos de comparación de cadenas para hacer las comparaciones entre los textos y las fórmulas y así encontrar las que están contenidas en ellos.

- La independencia del dominio es una característica propia de la teoría formulaica a diferencia de todos los métodos clásicos.

- Aún no se sabe si en efecto la implementación de la teoría formulaica es mucho más eficiente en tiempos computacionales que cualquier otro método clásico, pero éste no es el objetivo, en realidad lo que se buscó fue hacer una implementación sencilla, flexible, mucho más intuitiva y verificar que en efecto tiene resultados aceptables en tiempo de ejecución, pero principalmente que las fórmulas obtenidas sean relevantes.

- Se trabajó sobre documentos no estructurados en el formato de texto simple .txt. Se contó con dos conjuntos de textos: uno para el entrenamiento del algoritmo y otro para las pruebas.

1.4. Limitaciones

- La teoría formulaica es una teoría lingüística pura, no hay aplicaciones computacionales ni formalización de terminología o conceptos.

- Sólo podemos usar dominios formulaicos, la teoría no aplica por ejemplo a textos literarios los cuales se caracterizan por evitar la utilización de frases "trilladas.º simples repeticiones.

- También se tomó en cuenta que debido a que se trata de una teoría para el idioma inglés, podra generar algunos problemas cuando se trata de otros idiomas

- No hay córpora recolectados de dominios formulaicos, así que esto podría complicar la realización de los experimentos.

- No existen trabajos relacionados, es por ello que el carácter de nuestra investigación será exploratorio; al no haber experimentos similares no tendremos una guía y por lo tanto será difícil tener garantías de resultados satisfactorios.

- Se exploró la opción de SUBDUE como posible herramienta para reconocer patrones y no hay precedente de que haya sido utilizado en un estudio como éste.

- Existen algunos problemas cuando hay estribillos, o estructuras similares, en los textos porque hay que encontrar el tamaño de ideal que no divida el estribillo en varias partes y eso nos genere varias fórmulas cuando en realidad debería de ser una sola. Esto se trabajará encontrando el tamaño máximo de una oración dentro del texto, y así se evitarían estos problemas, aunque esto nos hará enfrentar el problema de segmentación de textos en oraciones. Se pretende utilizar, por ejemplo, el punto como símbolo de fin de oración, aunque no será tan sencillo porque existen excepciones como en el caso de las abreviaturas que llevan punto pero no indican fin de oración; no sabemos si esto nos generará otros problemas.

- Existen ciertas dificultades con las fórmulas que tienen variables, hay que marcar diferencia entre las variables que pueden ser vacías y las que forzosamente deben tomar el valor de una palabra. En realidad es un problema de representación, para ello se revisarán algoritmos que nos puedan servir para comparar cadenas de tamaños distintos y determinar si son parecidas aunque tengan variables como parte de la fórmula. Además es importante analizar cuál será el número máximo ideal de variables que pueden estar contenidas en una fórmula, así como el máximo de variables consecutivas.

- El éxito de la fase de clasificación depende mucho de si se tiene éxito en la fase de entrenamiento, es decir, dependerá de si se logra construir una lista de fórmulas realmente buena.

1.5. Organización del documento

La definición de la teoría formulaica y su terminología es presentada en el capítulo 2; algunas áreas y trabajos relacionados también se incluyen en el capítulo 2. En éste también se exploran algunos algoritmos que nos podrían ser útiles en la etapa de implementación.

Después en el capítulo 3 se presenta la formalización del sistema. En éste se incluyeron definiciones matemáticas de algunos conceptos (términos) presentados en el capítulo 2.

Posteriormente en el capítulo 4 se presentó el diseño del sistema, que incluye la arquitectura, algunos casos de uso y la descripción de los módulos.

Basados en el diseño se presenta la implementación del sistema en el capítulo 5, ilustrándose con diagramas de secuencias, interfaces y describiendo las principales funcionalidades del sistema.

Las pruebas realizadas y la discusión de los resultados son presentadas en el capítulo 6.

En el capítulo 7 se presentan sugerencias para continuar este proyecto.

Por último las conclusiones en el capítulo 8.