XXV Jornadas de Automática
Ciudad Real, del 8 al 10 de septiembre de 2004

CLASIFICACIÓN TEXTUAL BASADA EN TÉRMINOS JERÁRQUICOS

Francisco Javier Panizo, José R. Villar, Ángel Alonso
Área de Ingeniería de Sistemas y Automática, Dpto. Ingeniería Eléctrica y Electrónica, Universidad de León, Campus de Vegazana S/N 24071 León España
inffjp00@estudiantes.unileon.es, {diejvf, dieaaa}@unileon.es

José M. Alija
Área de Lenguajes y Sistemas Informáticos, Dpto. Dirección y Economía de la Empresa, Universidad de León, Campus de Vegazana S/N 24071 León España
ddejap@unileon.es

Resumen

Ante la inmensa cantidad de documentos manejados hoy en día por cualquier organización se usan clasificadores como solución idónea, ya que pretendemos automatizar. El grupo de investigación está desarrollando una base de conocimiento en ingeniería de control, base de conocimientos que se utilizará en el diseño de sistemas soporte en el diseño de controladores. Para tal fin, y con el objetivo de facilitar la educción y evolución de la base de conocimientos, se planteó implementar técnicas para la generación automática bien de ontologías, instancias de la base de conocimientos, o reglas del dominio. El presente artículo muestra el trabajo realizado para la consecución del clasificador de textos necesario para implementar las técnicas antes mencionadas.

1 Introducción

Ante la inmensa cantidad de documentos manejados hoy en día por cualquier organización, que necesitan de una búsqueda y consulta rápida y fiable, se usan entonces, clasificadores como solución idónea, ya que pretendemos automatizar, total o parcialmente, el tedioso y lento proceso de clasificar un documento dentro de una serie de términos, para poder realizar búsquedas temáticas, pues este es el principio básico de cualquier clasificador, relacionar un documento con una serie de términos bien conocidos.

Existen muy variadas técnicas de clasificación, principalmente basadas en inteligencia artificial y aprendizaje automático, y dentro de esta categoría, bien por procesamiento de lenguaje natural y determinación de la vaguedad del significado de las palabras, bien por métodos basados en entrenamiento probabilístico, bien por métodos basados en árboles de decisión, bien por métodos basados en la extracción de información, bien por métodos basados en reglas, por métodos basados en ejemplos, por redes neuronales, etc.

Una de las técnicas posibles es la de extracción de información, que se basa en el procesamiento de los documentos a tratar, con el fin de provocar una educción de términos representativos presentes en estos textos. Así pues, esta aproximación será la usada si deseamos desarrollar un sistema clasificador con un entrenador autónomo, con el consiguiente ahorro en tiempo y esfuerzo en la creación de un corpus. Sin embargo esta solución también presenta ciertos inconvenientes, como la elección adecuada de un conjunto de documentos representativos de un término específico, además de la posibilidad de realizar aproximaciones erróneas a ciertos términos.

El grupo de investigación está desarrollando una base de conocimiento en ingeniería de control, base de conocimientos que se utilizará en el diseño de sistemas soporte en el diseño de controladores. Para tal fin, y con el objetivo de facilitar la educción y evolución de la base de conocimientos, se planteó implementar técnicas para la generación automática bien de ontologías, instancias de la base de conocimientos, o reglas del dominio ([3], [4], [5] y [6]). Tras un estudio de dichas técnicas se decidió implementar el algoritmo de clasificación por palabras simples, modificado, para luego implementar la extracción de información [6].

El presente artículo muestra el trabajo realizado para la consecución del clasificador de textos necesario para implementar las técnicas antes mencionadas. En la sección 1 se detallan las técnicas desarrolladas por Riloff, mostrando la adaptación realizada en este trabajo en la seccion 2. En la siguiente sección se muestran los ensayos y resultados. Finalmente, en la sección 4 se muestran conclusiones y trabajos futuros.

2 Algoritmo de clasificación por palabras simples de Riloff

El trabajo propuesto para la consecución de las metas propuestas, se basa en la utilización del concepto de clasificación textual para la resolución de problemas de ordenación y distinción de escritos.

El concepto de clasificación textual, es en si mismo, uno de los muchos problemas planteados con relación al tratamiento semántico de la información. Algunos de estos retos son tan ambiciosos como la creación de una red semántica, el filtrado de información, la extracción de información… que permitirían alcanzar objetivos tan interesantes como las búsquedas eficientes, basadas en conocimiento y no en palabras como hasta ahora. Para ello es necesario la constitución de una base de conocimientos u ontologías como principal soporte de desarrollo. Esta base sería la que nos permitiría realizar una clasificación de un documento en función a la información que nos transmite.

Comentar que algunos de estos problemas planteados tienen gran relación con nuestro problema principal, la clasificación textual. Así por ejemplo, se pueden usar las soluciones aportadas respecto al concepto de extracción de información como base para el desarrollo práctico de una solución de nuestro problema. Ello es así porque la extracción de información nos proporciona los términos necesarios para clasificar nuestros textos. Esto es especialmente importante, en tanto que este concepto de extracción de información, realiza un tratamiento del lenguaje natural, tomando como base para ello el valor contextual de ciertas palabras.

Para lograr el objetivo final de realizar un clasificador textual, se afrontarán distintos enfoques como usar herramientas totalmente autónomas, o por el contrario, otras supervisadas parcialmente, de manera que sea un experto quien realice un posterior análisis de los resultados obtenidos de forma automática. Este último enfoque sería de utilidad en tanto que, tanto el documento queda clasificado según el criterio último de un experto, como por el hecho de que el sistema podría realimentarse con la opinión del experto, de manera que futuras clasificaciones sean más precisas.

En función de nuestra solución práctica, como se comentó anteriormente, estos enfoques serán útiles y reaprovechables o no tendrán ningún sentido. Así por ejemplo, si adoptamos una solución con capacidad autónoma de aprendizaje, garantizamos una mejor escalabilidad desde nuestro dominio de actuación a cualquier otro, con un ligero esfuerzo.

Uno de los planteamientos estudiados corresponde al de Riloff [6], donde propone un algoritmo de palabras simples a la hora de realizar la clasificación textual. Así pues, se expone el hecho de que este algoritmo es similar al de la relevancia de claves, menos por el hecho de usar palabras en vez de dichas claves o conjuntos de palabras.

Hay que tener en cuenta que a la hora de crear un corpus, este se crea únicamente con información acerca de las palabras, es decir, no utiliza términos para clasificar un determinado documento, sino que lo hace relevante dentro de un determinado contexto, siempre y cuando contenga alguna de las palabras relevantes que se han calculado para dicho contexto, representado por un conjunto de documentos relevantes del mismo.

Dicho contexto se representa por un conjunto de documentos, tanto relevantes como no dentro de él. Así pues, el paso previo antes de obtener un clasificador basado en palabras simples, sería del entrenarlo mediante dichos documentos de la manera que se detalla: para cada palabra que aparezca en el conjunto de entrenamiento, se procede a contar el número de veces que aparece en dicho conjunto así como el número de veces que aparece en un documento relevante del mismo. Así pues, con estos datos, se estima la probabilidad condicional de que un texto sea relevante, teniendo en cuenta que contiene esa palabra. La fórmula es como sigue:

$$Pr = Npali / Npali€rel$$

Fórmula1:Probabildad condicional

En ésta Npali es el número de veces que la palabra iesima aparece en el conjunto de entrenamiento. Npali€rel es el número de veces que dicha palabra aparece en documentos relevantes del conjunto de entrenamiento.

Una vez hallado este valor, se usan dos valores umbrales, R y M para seleccionar el conjunto de palabras que influyen en la relevancia de un cierto documento. Así pues, consideramos que una palabra es relevante si su probabilidad condicional es mayor o igual a R y si aparece al menos M veces en el conjunto de entrenamiento.

Por último, clasificamos un cierto documento como relevante para el contexto que se está tratando, si contiene alguna de dichas palabras releventes.

3 Algoritmo de clasificación por términos jerarquizados

Una vez planteadas las soluciones antes mencionadas, se optó por el desarrollo de una solución basada en un algoritmo de palabras simples modificado, o lo que es lo mismo, clasificar un documento dentro de un cierto conjunto de términos, en función de las palabras que lo componen, sin tener en cuenta otros factores como el papel semántico que ocupan dentro de la parte en proceso de estudio, idea en la que se basa el concepto de extracción de información.

Así pues, a la hora de clasificar, la primera tarea a realizar es la de determinar, tanto las palabras que aparecen en la descripción de un documento (a la postre, la forma de clasificar), como su número, almacenando esta información en la base de datos que nos sirve de soporte. Hay que resaltar el hecho de que en esta parte del proceso, se procede a un filtrado de todas aquellas palabras que no aportan ninguna información relevante, como determinantes, artículos, preposiciones...

El algoritmo desarrollado, basado en el propio de Riloff de palabras simples, y modificado para nuestro caso, es el siguiente:

Nos cercioramos de que el documento no se encuentre ya clasificado
Obtenemos las palabras de la descripción, así como el número ocurrencias
Obtenemos los datos almacenados de cada término
Para cada palabras relacionadas con cada termino
Buscamos la igualdad con las de la descripción
Si la palabra del término se encuentra en la descripción
Multiplicamos el número de veces que aparece dicha palabra por su certidumbre y lo sumamos al totalTerm para dicho documento
Ordenamos los vectores necesarios con los cálculos para los términos
Insertamos la relación resultante entre el documento y el término para los k-esimos mejores
Si no hay k-esimos términos con los que clasificar
Clasificamos con los que tengamos disponibles
Si ocurre algún error a la hora de guardar la clasificación
Mensaje de error al usuario
Borramos todas las posibles clasificaciones

Partimos del hecho de poseer un corpus semántico que sin duda es el corazón de nuestra aproximación a un primer clasificador. Hay que tener en cuenta el hecho de que el corpus semántico en si mismo es un trabajo concienzudo y pormenorizado en el que se utiliza más tiempo que en el propio desarrollo de la herramienta clasificadora. Recordar que en el caso del algoritmo de palabras simples, este corpus se crea de manera automática, tomando como base una serie de documentos relevantes, lo cual facilita mucho esta labor.

Este corpus es en sí mismo el que nos da la información de términos, sus relaciones, así como de las palabras que definen a un término, por lo que supondrá la base sobre la que se desarrollará nuestro clasificador.

Se implementó una herramienta para introducir el corpus semántico para cada término. Se necesita disponer de una archivo de texto formateado según la siguiente estructura

D:Término;TérminoPadre;RT:<TérminosRelacionados>*;PALABRAS: palabra certidumbre <, palabra certidumbre>*;

Esta estructura indica cual es el *Término* así como cual es el *TérminoPadre*, con cuales está relacionado, *TérminosRelacionados*, y su corpus semántico, que está formado por las palabras indicadas, cada palabra con su certidumbre acerca de la incumbencia sobre el significado del término. Un ejemplo de una línea de dicho archivo sería:

D:BIOLOGIA;MEDIO NATURAL;RT:MEDIO NATURAL,BIOTECNOLOGIA;PALABRAS: BIOLOGÍA 0.5, BIOQUÍMICA 0.1, BOTÁNICA 0.1, CITOLOGÍA 0.1, BIOGENÉTICA 0.3;

Contando con esto, se diseñó una herramienta que almacenara en el correspondiente medio de soporte toda esta información, con el fin último de utilizarla para resolver nuestro problema de clasificación textual. Se llegó a este compromiso después de un análisis pormenorizado, tomando dicha aproximación, como base para el desarrollo futuro de nuevas versiones con mejoras sustanciales a la hora de educir los términos que representan la clasificación. Sin embargo este hecho no empaña la solución adoptada, pues ésta cuenta con una gran relación de eficacia a la hora de clasificar, si bien es cierto, depende mucho de dicho fichero de corpus, con lo que la escalabilidad resulta uno de sus puntos débiles. Además el hecho de clasificar basándonos en palabras, nos lleva al hecho de errar, pues no se cuenta, por ejemplo, con un tratamiento de sinónimos.

Una vez creado el corpus que nos servirá de base, se procedió al diseño de la herramienta clasificadora. El trabajo que realiza ésta es pues de gran utilidad, y es éste un cierto esquema de clasificación que es ampliable a otras soluciones basadas en palabras simples.

La primera tarea que realiza es recoger toda la información necesaria almacenada sobre el corpus.

Después se procede a la obtención de la información ya filtrada y procesada acerca de un cierto documento, de la cual, principalmente, nos interesa saber el número de palabras totales del documento, así como, la frecuencia de una determinada palabra dentro del mismo.

A continuación se procede a realizar el conteo de palabras, que perteneciendo al documento, se encuentran dentro del corpus. Si se produce esta situación, se calcula la certidumbre acumulada para los términos que contienen dicha palabra, contando al finalizar el proceso, con un cierto grado de correspondencia entre el documento en cuestión y todos los términos con los que contamos en el corpus.

Posteriormente realizamos la clasificación propiamente dicha, asociando un conjunto de términos al documento, siempre y cuando los términos superen un cierto umbral de correspondencia, que podemos indicarle, eligiendo los k-esimos mejores términos, o bien, por un valor umbral numérico, tanto de palabras coincidentes, como de certidumbre.

4 Ensayos y pruebas

4.1 Tesauro y corpus: descripción

A la hora de realizar las pruebas sobre el clasificador, lo primero a tener en cuenta, es saber que tipo de tesauro se ha usado para la clasificación. Esto es así, porque tal y como se explicó anteriormente, esta parte, nos proporcionará los términos con los que vamos a clasificar un determinado documento.

Así pues, el tesauro utilizado trata del campo de la biología, por lo que, las palabras que componen su corpus también lo son. Se trata de un tesauro pues que cuenta con 193 términos, y de 108 relaciones entre un término y su termino padre. Además se definen 111 relaciones entre términos hermanos, es decir, sin dependencia jerarquica vertical.

4.2 Descripción de pruebas

A la hora de realizar las pruebas se procedió a elegir tres términos del tesauro, con el fin de encontrar documentos que trataran de estos temas, y que sirvirán para relizar nuestras pruebas con un cierto grado de seguridad. Así se eligieron los términos: mamífero, carnívoro y herbívoro para dichas pruebas. La elección de dichos términos no fue aleatoria, sino guiada, ya que entre el primero y los dos siguientes existe una relación de paternidad jerarquica. Así pues, con la elección de estos tres términos, se pretendía demostrar el hecho de que el clasificador no solo realizaba una clasificación muy guiada, sino que se pretendía demostrar su alto grado de precisión a la hora de discernir entre términos muy relacionados entre sí.

Las pruebas se realizaron en dos series con 22 documentos para la primera y 24 para la segunda, de manera que la primera de ellas contaba, en su totalidad, con documentos con un número suficiente de palabras. La segunda de las series contenía otros 24 documentos con un número de palabras del doble, en terminos de media, que los documentos que formaban la primera serie. El número de términos con los que clasificar un cierto documento, se fijó en un máximo de 2 términos por documento.

4.3 Resultados obtenidos

Explicar que los datos obtenidos han sido clasificados en TP (verdadero positivo), FP (falso positivo), TN (verdadero negativo) y FN (falso negativo)

Respecto a la primera serie de documentos tratada, para los tres términos con los que contamos, se han obtenido los siguientes datos:

	TP	FP	TN	FN
Mamífero	5	3	7	7
Carnívoro	6	2	1	13
Herbívoro	2	4	0	16

Tabla1: Resultados de la primera serie

Para la segunda serie de documentos clasificados, se han obtenido los siguientes resultados:

	TP	FP	TN	FN
Mamífero	6	2	4	12
Carnívoro	3	5	2	14
Herbívoro	3	5	1	15

Tabla2:Resultados de la segunda serie

5 Conclusiones y trabajos futuros

La conclusión principal es el hecho de que un algoritmo como el implementado para la consecución de los objetivos marcados, a pesar de sus problemas (escalabilidad y ampliación), resulta de gran utilidad en entornos acotados y pequeños, además de bastante precisa para ciertos términos con un gran parecido o con una relación visible entre ambos, por lo que se demuestra que no siempre la solución mejor es la más ambiciosa. Sin embargo, en caso de haber dispuesto de un conjunto de pruebas mayor se habría demostrado la mejora de los niveles obtenidos.

Además se podría plantear el reto de realizar una sustancial mejora en el motor clasificador con el objeto de poder implementar un cierto nivel de inteligencia al sistema a la hora de realizar la clasificación, por ejemplo, permitiendo que el sistema clasifique únicamente a un texto dentro de su término más específico, y no dentro de otro superior jerárquicamente, si bien se debería dejar la posibilidad de que un experto realice una comprobación, por el hecho de que se quiera clasificar realmente en el término más genérico.

Como trabajos futuros cabe destacar proveer de un algoritmo que realice el entrenamiento automático, desarrollando para ello una solución basada en el concepto de extracción de información. Esta mejora sin embargo, no modificaría en gran medida el motor clasificador utilizado por el momento y explicado en este artículo, siempre y cuando la extructura del corpus que se extraiga del análisis de ficheros de entrenamiento mantega ciertos requisitos de diseño que cumple el actual. También, como trabajo futuro, incluir el presenta trabajo en un sistema para ayuda en la generación de propuestas conceptuales para el crecimiento y evolución de ontologías.

Referencias

[1] Attardi, G.; Di Marco, S.; Salvi, D.; Sebastiani, F. (1998) Categorisation by Context,Journal of Universal Computer Science

[2] Riloff, E. (1991) Little Words Can Make a Big Difference for Text Classification, Proceedings of the 18th Annual International ACM SIgIR Conference on Research and Development in Information Retrieval

[3] Riloff, E. (1996) Using learned extraction patterns for text classification, Connectionist, Statistical, and Symbolic Approaches to Learning for Natural Language Processing. In Wermter, S., Riloff, E., & Scheler, G. (eds.)., Springer-Verlag

[4] Riloff, E. (1993) Automatically Constructing a Dictionary for Information Extraction Task, Proceedings of the Eleventh National Conference on Artificial Intelligence, AAI Press – MIT Press

[5] Riloff, E. (1996) An Empirical Study of Automated Dictionary Construction for Information Extraction in Three Domains, AI Journal

[6] Riloff, E.; Lehnert, W. (1994) Information Extraction as a Basis for a High-Precision Text Classification, ACM transactions on Information Systems