Clasificación Automática de Textos de Desastres Naturales en México

Alberto Téllez-Valero, Manuel Montes-y-Gómez,
Olac Fuentes-Chávez, Luis Villaseñor-Pineda
Instituto Nacional de Astrofísica, Óptica y Electrónica (*INAOE*)
Luis Enrique Erro No. 1, Sta. María Tonanzintla, C.P. 72840 Puebla, México
albertotellezv@ccc.inaoep.mx, {mmontesg, fuentes, villasen}@inaoep.mx

Resumen

La búsqueda de información de dominio específico en sitios web es una tarea compleja. Actualmente, las máquinas de búsqueda facilitan la localización de información en Internet, sin embargo, por su principio de generalidad, y por la ambigüedad del lenguaje, éstas presentan bajos niveles de precisión. En este trabajo se demuestra que el filtrado de páginas web con información relevante a un dominio específico se puede realizar satisfactoriamente usando técnicas de clasificación de texto. Se presentan algunos experimentos en la clasificación automática de noticias de desastres naturales en México. Los resultados obtenidos indican que es posible clasificar una página web dentro de las categorías de huracán, inundación, sequía y no relevante con una exactitud aproximada del 97%.

Palabras clave: clasificación automática de textos, recuperación de información, aprendizaje automático.

1. Introducción

La web es un medio que permite acceder una gran variedad de información. Sin embargo, su alta heterogeneidad (i.e. información con distintas temáticas, estilos, formatos, e idiomas), hace día a día más difícil encontrar la información relevante, y en consecuencia, extraer el conocimiento potencialmente útil para la toma de decisiones.

Los métodos y tecnologías de búsqueda de información en Internet solucionan parcialmente este problema de acceso a la información [1,2]. Ellos son suficientemente generales para auxiliar en todo tipo de consultas, pero son inapropiados para buscar información muy específica. Por su parte, los métodos de clasificación de texto [3] son eficientes para filtrar información relevante para un dominio específico o tema en cuestión (para el cual fueron entrenados), pero no tienen la funcionalidad necesaria para acceder a información distribuida por toda la web.

Actualmente en el Laboratorio de Tecnologías del Lenguaje del INAOE se está diseñando un sistema que combina ambos tipos de métodos, de búsqueda de información y de clasificación de textos, con el fin de crear automáticamente un repositorio de información de desastres naturales en México[1]. La disponibilidad de este almacén de información permitirá adquirir un mejor conocimiento sobre los fenómenos naturales desastrosos, y con ello aprender a prevenir y minimizar sus efectos.

En este artículo presentamos un experimento relacionado con el filtrado de noticias relevantes al tema de desastres naturales. Este experimento considera varias estrategias de extracción de características de los documentos, así como diferentes métodos de clasificación de texto. En particular se emplean las estrategias de reducción de dimensionalidad conocidas como umbral en la frecuencia y ganancia en la información, además de los métodos de clasificación simple de Bayes y vecinos más cercanos.

El resto del documento se organiza de la siguiente manera. La sección 2 describe el conjunto de entrenamiento usado en la construcción del clasificador. La sección 3 expone los pasos empleados para transformar el texto, inicialmente en formato HTML, a una representación adecuada para la tarea de clasificación. La sección 4 introduce los métodos de clasificación empleados, en este caso el simple de Bayes y el de vecinos más cercanos. Finalmente, la sección 5 presenta los resultados obtenidos y la sección 6 expone nuestras conclusiones y el trabajo a futuro.

[1] Recolección, Extracción, Búsqueda y Análisis de Información en Español. Proyecto CONACYT U39957-Y.

2. Conjunto de entrenamiento

La construcción de un clasificador de textos mediante métodos de aprendizaje automático requiere de un conjunto de documentos previamente clasificados, conocido como conjunto de entrenamiento. En nuestro caso, usamos el periódico Reforma (www.reforma.com) como fuente de información principal. De este sitio se recopilaron noticias relacionadas (tanto relevantes como irrelevantes) con los fenómenos naturales de huracán, inundación y sequía correspondientes a los últimos dos años. Las noticias relevantes incluyen información del fenómeno natural, mientras que las catalogadas como irrelevantes contienen palabras o frases usadas comúnmente en la descripción de un fenómeno natural pero que se usan en contextos muy diferentes. Por ejemplo, la palabra huracán en el contexto de "el presidente está en el ojo del huracán".

El conjunto de entrenamiento obtenido finalmente consistió de 375 documentos, de los cuales el 11.5 % son noticias relevantes y el 88.5 % restante son irrelevantes. Estas cifras son un reflejo de la dificultad para recuperar información de desastres naturales en Internet usando las técnicas actuales de búsqueda, pero a su vez son un reflejo de la distribución real de las noticias que contienen palabras relacionadas con la descripción de estos fenómenos naturales (que en su mayoría es información irrelevante).

3. Extracción de características

Una vez obtenido el conjunto de documentos de entrenamiento, el siguiente paso en la construcción de un clasificador de textos consiste en transformar los documentos, de su formato inicial, a una representación adecuada para el algoritmo de aprendizaje, y en general para la tarea de clasificación. Básicamente, esta parte del proceso considera la extracción de las características principales (conjunto de palabras) de los documentos del conjunto de entrenamiento. A continuación se describen las diferentes etapas empleadas en la extracción de dichas características.

3.1 Pre-procesamiento

El propósito de esta etapa es reducir el tamaño de los documentos eliminando las partes de los textos que no son relevantes, es decir, que no dicen nada sobre su contenido. El proceso realizado a cada uno de los documentos fue el siguiente:

- *Eliminación de etiquetas HTML* – debido a que los documentos son páginas Web y las etiquetas HTML no proporcionan información útil en nuestra tarea de clasificación.
- *Eliminación de símbolos de puntuación* – como el análisis no es del tipo semántico, estos elementos no son necesarios.
- *Eliminación de palabras vacías* – palabras frecuentes que no transmiten información (por ejemplo pronombres, preposiciones, conjunciones, etc).
- *Reducir palabras a su raíz* - eliminar sufijos y afijos de una palabra de tal modo que aparezca sólo su raíz léxica (por ejemplo desconocer, desconocerlos y desconocía tienen la raíz léxica desconoc).

La reducción en tamaño de cada documento fue en promedio aproximadamente del 52 % de su tamaño original.

3.2 Indexado

La representación de los documentos más comúnmente usada es el llamado modelo vectorial [4]. En este modelo, los documentos son representados por vectores de palabras en un espacio de dimensionalidad n, siendo n el número de palabras diferentes de los documentos (el vocabulario).

En el modelo vectorial, la colección de documentos se representa por una matriz A de palabras por documentos, donde cada entrada representa el peso de una palabra en un documento. Esto es, $A = (a_{ik})$, donde a_{ik} es el peso de la palabra i en el documento k.

Existen varias maneras de determinar el valor de a_{ik}, en este experimento se uso el ponderado booleano. Este tipo de indexado (ponderado) es la aproximación más simple y por lo tanto más rápida de calcular. La idea consiste en asignar 1 si la palabra i ocurre en el documento k y 0 en caso contrario.

En el proceso de indexado se encontró que el conjunto de entrenamiento preprocesado contenía 143,961 instancias léxicas con un vocabulario de 14,562 palabras diferentes. Por lo tanto, el tamaño de la matriz *A* necesaria para representar el conjunto de documentos de entrenamiento es de dimensiones 375×14,562. Además, las frecuencias de las palabras en el vocabulario varían entre 1 y 977 en la colección total de documentos.

Los términos más frecuentes en el conjunto de entrenamiento fueron: México (977), año (754), huracán (575), agua (401) y gobierno (452).

3.3 Reducción de dimensionalidad

A partir de los datos mostrados en la sección anterior se advierte uno de los problemas centrales en la clasificación de texto, la alta dimensionalidad del espacio de características (375×14,562 en nuestro caso). Las técnicas de clasificación estándar no pueden tratar con tales conjuntos de características, puesto que el procesamiento es extremadamente costoso en términos computacionales. Además, los resultados llegan a ser poco confiables debido a las deficiencias en los datos de entrenamiento al tomar en cuenta palabras con poca información del dominio, por ejemplo, palabras con una frecuencia igual a uno en toda la colección.

Así pues, debido a los problemas causados por la alta dimensionalidad de la representación del conjunto de entrenamiento, existe la necesidad de reducir el conjunto de características original, es decir, hacer una reducción de dimensionalidad. En nuestros experimentos, usamos el umbral en la frecuencia y la ganancia en la información (IG) [8] como mecanismos principales de esta reducción.

El umbral en la frecuencia calcula la frecuencia para cada palabra en el corpus de entrenamiento y elimina las palabras donde su frecuencia fue menor que el umbral predeterminado. Es decir, la idea básica es que las palabras raras no proporcionan información para predecir la categoría. Por su parte, la ganancia de información mide el número de bits de información obtenida para predecir la categoría por medio de la presencia o ausencia de una palabra en el documento. La elección de estos métodos de reducción se debe a que evaluaciones presentadas han revelado que estos métodos se encuentran entre los más efectivos [5].

Ahora bien, para reducir la dimensionalidad de la matriz *A* se realizaron experimentos utilizando en primer lugar un umbral en la frecuencia y posteriormente ganancia en la información, para esto se calcula la frecuencia y la ganancia en la información para cada una de las palabras en el vocabulario. Posteriormente, en el primer experimento se eliminaron aquellos términos cuya frecuencia fue menor a 10, el resultado fue una dimensión de 375×2550 para nuestro espacio de características. En el experimento posterior se eliminaron aquellos términos cuya ganancia en la información fue cero. En otras palabras, sólo se tomaron aquellos términos que dan información útil para predecir la clase. El resultado fue una nueva dimensión de 375×214 para nuestro espacio de características, lo que refleja que sólo el 0.7 % del vocabulario es información útil para predecir la clase. Los términos del vocabulario con mayor ganancia en la información fueron: meteorología (0.1327), tropical (0.1215), sequía (0.1105), viento (0.0974) y agua (0.0942).

4. Método de clasificación

Un número importante de técnicas de clasificación estadística y aprendizaje automático se han empleado en la clasificación de textos, por ejemplo: vecinos más cercanos [5], árboles de decisión [6], clasificador simple de Bayes [6], máquinas de vectores de soporte [7] y Rocchio [8], entre otros.

Con base en los resultados reportados en la bibliografía reciente [8][10], se seleccionaron los métodos tradicionales de vecinos más cercanos y clasificador simple de Bayes para nuestro experimento. Vecinos más cercanos es un método basado en instancias, donde no se construye ninguna descripción de las categorías, más bien se memorizan directamente los juicios de relevancia de los usuarios (conjunto de entrenamiento) para la clasificación de ejemplos no vistos. Por su parte, el método simple de Bayes es de tipo probabilístico, en él se usa el conjunto de entrenamiento para estimar los parámetros de una distribución de probabilidad que describa el conjunto de entrenamiento. Ambos algoritmos han mostrado ser de los mejores en la tarea de clasificación de textos.

5. Resultados

A continuación se muestran los resultados obtenidos en los experimentos. Estos resultados consideran los métodos de vecinos más cercanos y clasificador simple de Bayes. Asimismo analizan el efecto de aplicar reducción de dimensionalidad con las técnicas de umbral en la frecuencia y ganancia en la información al conjunto de entrenamiento.

La evaluación de la efectividad de los clasificadores se basó en el método de validación cruzada con 10 pliegues (10 Fold Cross Validation, en inglés) [11][12]. En la tabla 5.1 se presentan los porcentajes de aciertos y fallos de cada clasificador, además del error cuadrático medio (RMS). Cabe señalar que estas medidas son comunes en el área de aprendizaje automático.

	Umbral en la frecuencia Frec > 10		**Ganancia en la información IG > 0**	
	Vecinos más cercanos (K=1)	**Simple de Bayes**	**Vecinos más cercanos (K=1)**	**Simple de Bayes**
Instancias clasificadas correctamente	90.93 %	93.3 %	92.8 %	97.06 %
Instancias clasificadas incorrectamente	9.06 %	6.6 %	7.2 %	2.93 %
Error cuadrático medio	0.2117	0.182	0.1833	0.1195

Tabla 5.1 Evaluación con mediadas de aprendizaje automático.

En la tabla 5.2 se presenta la matriz de confusión del clasificador simple de Bayes que utiliza ganancia en la información para reducir la dimensionalidad. En esta matriz, la diagonal principal refleja las instancias clasificadas correctamente, y los valores fuera de la diagonal representan los errores cometidos por el clasificador.

	Huracán	**Inundación**	**Sequía**	**No relevante**
Huracán	15	0	0	3
Inundación	0	5	0	2
Sequía	0	0	14	4
No relevante	1	0	1	330

Tabla 5.2 Matriz de confusión.

Otra manera de evaluar un clasificador de textos es aplicando medidas provenientes de la recuperación de información. Estas medidas son la precisión, el recuerdo, y la medida-F [9]; siendo esta última una combinación lineal de las otras dos. Las tablas 5.3 y 5.4 muestran los resultados obtenidos con ambos clasificadores.

Umbral en la frecuencia Frec > 10			**Ganancia en la información IG > 0**			
Precisión	**Recuerdo**	**Medida -F**	**Precisión**	**Recuerdo**	**Medida -F**	**Clase**
0.889	0.444	0.593	0.846	0.611	0.71	Huracán
1	0.143	0.25	1	0.714	0.833	Inundación
0.667	0.111	0.19	0.6	0.333	0.429	Sequía
0.912	0.994	0.951	0.939	0.982	0.96	No relevante

Tabla 5.3 Evaluación de vecinos más cercanos.

Umbral en la frecuencia Frec > 10			**Ganancia en la información IG > 0**			
Precisión	**Recuerdo**	**Medida -F**	**Precisión**	**Recuerdo**	**Medida -F**	**Clase**
1	0.278	0.435	0.938	0.833	0.882	Huracán
0	0	0	1	0.714	0.833	Inundación
0.933	0.778	0.848	0.933	0.778	0.848	Sequía
0.932	0.997	0.964	0.973	0.994	0.984	No relevante

Tabla 5.4 Evaluación de simple de Bayes.

6. Conclusiones y trabajo a futuro

El mejor resultado obtenido en la clasificación de noticias de desastres naturales fue de 97 % de exactitud. Este resultado se logró aplicando el algoritmo simple de Bayes sobre el conjunto de entrenamiento reducido en su dimensionalidad mediante la técnica de ganancia de información.

Dos puntos importantes a resaltar de los experimentos son, en primer lugar, que se obtuvieron resultados aceptables en la clasificación automática de noticias de desastres naturales a pesar de la enorme diferencia entre el número de documentos relevantes e irrelevantes del conjunto de entrenamiento. En segundo lugar, que con ambos clasificadores los errores surgieron exclusivamente por clasificar documentos relevantes como no relevantes, y nunca por confundir documentos relevantes a una clase con documentos relevantes a otra. Esto último se refleja en una elevada medida de precisión, pero con un menor recuerdo (ver tabla 5.2, 5.3 y 5.4).

Los resultados obtenidos nos motivan porque muestran factible la construcción automática de un repositorio de información de desastres naturales en México.

Con respecto al trabajo futuro se planea ampliar el número de clases que maneja el clasificador actualmente, para lo cual estamos recolectando más documentos para el conjunto de entrenamiento. Además se considera construir un sistema que se adapte a los ejemplos de prueba mal clasificados; para ello se requiere de una mínima intervención del usuario. Por último, se continuará experimentando con otros métodos de representación de los documentos, así como con otros algoritmos de clasificación de texto.

Agradecimientos

El presente trabajo se realizó bajo el contexto del proyecto de investigación *Recolección, Extracción, Búsqueda y Análisis de Información en Español* financiado parcialmente por el CONACYT (Proyecto U39957-Y). Los autores también agradecen el apoyo del Laboratorio de Tecnologías del Lenguaje del INAOE.

Bibliografía

1. W. Hu, "World Wide Web Search Technologies", In Architectural Issues of Web-Enabled Electronic Business, Idea Group Publishing, 2002.
2. H. Chu and M. Rosenthal, "Search Engines for the World Wide Web: A Comparative Study and Evaluation Methodology", Proceedings of the ASIS 1996 Annual Meeting, 1996.
3. K. Aas and L. Eikvil, "Text Categorisation: a Survey", Technical Report, Norwegian Computing Center, 1999.
4. G. Salton and M. J. McGill, "An Introduction to Modern Information Retrieval," McGraw-Hill, 1983.
5. Y. Yang and J. P. Pedersen, "Feature Selection in Statistical Learning of Text Categorization", Proceedings of the 14th International Conference on Machine Learning, 1997.
6. D. Lewis and M. Ringuette, "A Comparison of two Learning Algorithms for Text Classification", Proceedings of the Third Annual Symposium on Document Analysis and Information Retrieval, 1994.
7. T. Joachims, "Text Categorization with Support Vector Machines: Learning with many relevant features", Proceedings of the 10th European Conference on Machine Learning (ECML), 1998.
8. F. Sebastiani, "Machine learning in automated text categorisation: a survey", Technical Report IEI-B4-31-1999, Istituto di Elaborazione dell'Informazione, 1999.
9. D. Lewis, "Evaluating Text Categorization", Proceedings of the Speech and Natural Language Workshop, 1991.
10. F. Sebastiani, "A Tutorial on Automated Text Categorisation", Proceedings of the 1st Argentinean Symposium on Artificial Intelligence (ASAI-99), 1999.
11. T. Mitchell, "Machine Learning", McGraw-Hill, 1997.
12. I. H. Witten and E. Frank, "Data Minning: Practical Machine Learning Tools and Techniques with Java Implementations", Morgan Kaufmann, 2000.