Extracción de Información Semántica a Partir de Categorías de Texto

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 las áreas de clasificación de texto y extracción de información enfocadas a la obtención de contenido semántico que permita la identificación de los documentos mediante descriptores o información asociada a las categorías.**

***Index Terms*— Document Classification, Document Clustering, Information Retrieval, Ontology Generation, 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, agruparlos en categorías que los relacionen en un contexto específico y tratar de 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 abordan conceptos importantes y enfoques establecidos para la obtención de información relevante a partir de documentos categorizados o agrupados, así como el almacenamiento y procesamiento de dicha información; posteriormente, en la sección III se consideran aspectos a tener en cuenta cuando se realizan tareas de procesamiento de texto, tales como la extracción de características importantes de dichos documentos. La sección IV muestra las técnicas mas usadas en el ámbito de la clasificación de documentos. Finalmente, en la sección V se presentan las conclusiones del estudio y se hace un análisis de las direcciones que debe tomar el desarrollo de técnicas que permitan solucionar el problema de manejo de la información.

II. Semántica de los Documentos

Se busca obtener información adicional de cada una de las categorías en las cuales se han clasificado los documentos, para ello se puede contar con un conocimiento previo acerca del dominio en el cual se encuentran las categorías de documentos que se están clasificando; este conocimiento puede ser extraido de ontologías públicas como en [19]. En [24] se presenta una técnica para dar explicaciones en el contexto de la clasificación, 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.

Se utilizan 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 [25].

Una vez se obtiene este conocimiento a parir de los documentos de las categor-ias determinadas, se deben encontrar técnicas que permitan administrarlo de forma que sea útil para las personas involucradas en el proceso, las etapas que intervienen en este proceso de administración del conocimiento conceptual son las siguientes [8]:

1. Obtención del conocimiento
2. Origanización y estructuración
3. Refinamiento
4. Distribución del conocimiento

II-A. Manejo de Información Mediante Ontologías

Una vez se han agrupado los documentos en las categorías correspondientes, 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.

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:

- Obtención del conocimiento
- Organización y estructuración del conocimiento
- Refinamiento del conocimiento
- Distribución del conocimiento

La idea final de este proceso es obtener soluciones inteligentes a problemas difíciles mediante la utilización adecuada de este conocimiento.

En [16] 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.

En [15] 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 le dan 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. Lo que se hace es tomar un conjunto de keyword generados por el usuario junto con la información contenida en el documento para ver 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 los keywords dados por el usuario y sus sinónimos. Posteriormente se aplica PCA para encontrar las palabras mas importantes.

III. Extracción de características

Para llevar a cabo una tarea de clasificación de texto, se debe contar primero con una repersentación adecuada de los documentos que se quieren clasificar.

Una de las principales dificultades en cuanto al manejo de este tipo de documentos, es la alta dimensionalidad debido a los esquemas que se utilizan para representar los documentos [1]. En este punto se aplican técnicas que permitan reducir la dimension de los documentos eliminando información que no es relevante para la tarea de clasificación.

Básicamente, la mayoría de técnicas para representar documentos se basan en un vector de palabras, en el cual cada posición representa una palabra, generalmente tomada de un diccionario construido previamente; la diferencia entre los diversos esquemas radica en el valor que se almacena para cada una de las palabras.

III-A. Palabras Características

Usualmente, la representación del conjunto de documentos se hace mediante una matriz A, donde cada entrada representa el peso de una palabra en un documento. [1]

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

en donde a_{ik} representa el peso de la palabra i en el documento k.

Para definir los diferentes esquemas de asignación de pesos, se definirán primero los siguientes valores: f_{ik} es la frecuencia de la palabra i en el documento k, N es el número de documentos, M es el número de palarbras en el diccionario y n_i es el número total de ocurrencias de la palabra i en todo el conjunto de documentos. Los esquemas más utilizados son los siguientes.

Boolean Weighting:

$$a_{ik} = \begin{cases} 1 & \text{si } f_{ik} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Word Frequency Weighting:

$$a_{ik} = f_{ik}$$

tf × idf Weighting:

$$a_{ik} = f_{ik} * \log \frac{N}{n_i}$$

tfc Weighting:

$$a_{ik} = \frac{f_{ik} * \log \frac{N}{n_i}}{\sqrt{\sum_{j=1}^{M} [f_{ik} * \log \frac{N}{n_i}]^2}}$$

ltc Weighting:

$$a_{ik} = \frac{\log(f_{ik}+1{,}0) * \log\frac{N}{n_i}}{\sqrt{\sum_{j=1}^{M}[\log(f_{ik}+1{,}0) * \log\frac{N}{n_i}]^2}}$$

Entropy Weighting:

$$a_{ik} = \log(f_{ik}+1{,}0) * (1 + \frac{1}{\log(N)}\sum_{j=1}^{N}[\frac{f_{ij}}{n_i}\log\frac{f_{ij}}{n_i}])$$

III-B. Otras Características

n-grams: Consiste en utilizar secuencias de palabras seleccionadas de la aplicación de técnicas estadísticas, por ejemplo, χ^2 [34]

Part-Of-Speech tagging: Consiste en asignar a cada palabra del documento una categoría sintáctica (verbos, adjetivos, etc). Se debe decidir para cada una de las palabras cual es la mejor categoría que la representa dentro del contexto del documento; puesto que una misma palabra puede tener diferentes categorizaciones de acuerdo a su utilización. [34]

Noun Phrases: Conocidas también como entidades, consiste en utilizar frases que en conjunto tienen un mismo significado, pero que individualmente no representan información relevante en el contexto, por ejemplo *sistema de televisión por cable* [34]. Se deben seleccionar solo aquellas frases que representen información importante para el proceso de clasificación; para realizar este proceso, se debe contar con inforación sintáctica de cada palabra como la dada por las técnicas anteriores [40].

Information Bottleneck: Esta técnica trata de obtener características de los documentos que se quieren clasificar de modo que dichas características seleccionadas no sean dependientes del contexto en el cual se llevará a cabo la tarea de clasificación [7].

Latent Semantic Analysis: Es un enfoque para indexar información contenida en documentos mapeando los documentos así como los términos en un espacio denominado *Latent Semantic Space*. Se parte de una representación de los documentos basada en la frecuencia de aparición de las palabras y se aplica una proyección lineal para reducir la dimensionalidad [23].

IV. Métodos de Clasificación

El problema de categorización de texto consiste en asignar automáticamente a cada documento una o varias de un conjunto predefinido de categorías. La mayoría de los enfoques propuestos en esta área trata de resolver el problema de clasificación binaria, es decir, para cada documento se debe determinar si pertenece o no a una categoría dada. En el caso de clasificación cuando se tienen más categorías, el problema se hace un poco más complejo puesto que un documento podría pertenecer a más de una categoría [1].

IV-A. Rocchio

Este método es el más utilizado tradicionalmente para tareas de clasificación. Consiste en construir un vector *prototipo* para cada una de las categorías; los documentos se clasifican calculando la distancia al vector prototipo. Los vectores prototipos para cada categoría se construyen como el vector promedio de todos los documentos de entrenamiento que pertenecen a dicha categoría [1].

IV-B. Clasificadores Bayesianos

En esta técnica se utilizan los datos de entrenamiento para estimar la probabilidad de cada categoría dados los valores de las características seleccionadas de un nuevo documento. Para estimar la probabilidad de tener una categoría dado un documento de prueba, se utiliza el teorema de bayes [1]:

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}$$

IV-C. K Nearest Neighbour

En esta técnica, cada vector de entrada es comparado con los demás documentos para encontrar los k documentos mas cercanos (o más similares) y utiliza las categorías de estos k documentos para determinar la categoría correspondiente del documento de prueba [1]. Una desventaja que presenta este algoritmo, es que se utilizan todas las características del documento para comparar distancias con todos los demás documentos, lo que lo hace lento en comparación con otras técnicas; se han presentado estrategias que tratan de resolver este problema encontrando pesos para cada característica del vector de modo que se puedan utilizar solo aquellas más representativas [22].

IV-D. Árboles de decisión

Los documentos de entrada son evaluados utilizando un árbol de decisión para determinar si son relevantes para el usuario o no (o si pertenecen a una determinada categoría), este árbol es construido a partir de los documentos de entrenamiento [1].

IV-E. Support Vector Machines

Esta técnica integra un componente de reducción de la dimensionalidad junto con la tarea de clasificación. Solo se aplica a problemas de clasificación binarios, de modo que los problemas con múltiples categorías deben ser tratados como una serie de problemas de clasificación binarios [1]. La idea es medir el margen de separación de los datos de diferentes categorías más que encontrar relaciones entre las características de los documentos y las categorías [6].

IV-F. Clasificación Jerárquica

En esta técnica, se tiene una jerarquía de categorías representada como un árbol, un documento puede ser asignado a una o mas de las categorías representadas por los nodos hoja del árbol [12].

IV-G. Otras Técnicas

Clasificación por reglas de decisión [4]. Clustering Incremental [11].

V. Conclusiones

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

[4] M.-L. Antonie and O. R. Zaïane. Text document categorization by term association. In *ICDM*, pages 19–26, 2002.

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

[6] A. Basu, C. R. Watters, and M. A. Shepherd. Support vector machines for text categorization. In *HICSS*, page 103, 2003.

[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] W. chiu Wong and A. W. chee Fu. Incremental document clustering for web page classification, Aug. 31 2000.

[12] S. DÁlessio, K. Murray, R. Schiaffino, and A. Kershenbaum. The effect of using hierarchical classifiers in text categorization. In *Proceeding of RIAO-00, 6th International Conference "Recherche dÍnformation Assistee par Ordinateur"*, pages 302–313, Paris, FR, 2000.

[13] F. Debole and F. Sebastiani. Supervised term weighting for automated text categorization. In *SAC*, pages 784–788, 2003.

[14] O. Drori. Identifying the subject of documents in digital libraries automatically using frequently-occurring words - study and findings, May 23 2003.

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

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

[17] 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.

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

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

[20] G. Guo, H. Wang, D. A. Bell, Y. Bi, and K. Greer. An kNN model-based approach and its application in text categorization. In A. F. Gelbukh, editor, *Proceedings of CICLING-04, 5th International Conference on Computational Linguistics and Intelligent Text Processing*, pages 559–570, Seoul, KO, 2004. Springer Verlag, Heidelberg, DE. Published in the "Lecture Notes in Computer Science" series, number 2945.

[21] E.-H. Han and G. Karypis. Fast supervised dimensionality reduction algorithm with applications to document categorization & retrieval. In *CIKM*, pages 12–19, 2000.

[22] E.-H. Han, G. Karypis, and V. Kumar. Text categorization using weight adjusted k-nearest neighbor classification. In *PAKDD*, pages 53–65, 2001.

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

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

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

[26] 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.

[27] G. Karypis and E.-H. Han. Concept indexing: A fast dimensionality reduction algorithm with applications to document retrieval and categorization. Computer science department TR-00-0016, University of Minnesota, 2000.

[28] T. Kudo and Y. Matsumoto. A boosting algorithm for classification of semi-structured text. In *Proceedings of EMNLP-04, 9th Conference on Empirical Methods in Natural Language Processing*, Barcelon, ES, 2004.

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

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

[31] B. Liu, Y. Dai, X. Li, W. S. Lee, and P. S. Yu. Building text classifiers using positive and unlabeled examples. In *ICDM*, pages 179–188, 2003.

[32] 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.

[33] R. E. Madsen, J. Larsen, and L. K. Hansen. Part-of-speech enhanced context recognition. In S. D. A.K. Barros, J. Principe, J. Larsen, T. Adali, editor, *Proceedings of IEEE Workshop on Machine Learning for Signal Processing XIV*, pages 635–644, Piscataway, New Jersey, Sept. 2004. IEEE Press.

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

[35] T. Oda and T. White. Developing an immunity to spam. In *GECCO*, pages 231–242, 2003.

[36] S.-B. Park and B.-T. Zhang. Co-trained support vector machines for large scale unstructured document classification using unlabeled data and syntactic information. *Information Processing and Management*, 40(3):421–439, 2004.

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

[38] D. Ramamonjisoa. Towards automated research topics discovery on scientific domain by agents system, Jan. 02 2003.

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

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

[41] M. Sinka and D. Corne. Evolving document features for web document clustering: A feasability study. In *Proceedings of the 2004 IEEE Congress on Evolutionary Computation*, pages 891–897, Portland, Oregon, 20-23 June 2004. IEEE Press.

[42] N. Slonim and N. Tishby. The power of word clusters for text classification. In *Proceedings of ECIR-01, 23rd European Colloquium on Information Retrieval Research*, Darmstadt, DE, 2001.

[43] A. Strehl, J. Ghosh, and R. Mooney. Impact of similarity measures on web-page clustering. In *Proceedings of the 17th National Conference on Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI 2000), 30–31 July 2000, Austin, Texas, USA*, pages 58–64. AAAI, July 2000.

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

[45] A. Sun and E.-P. Lim. Hierarchical text classification and evaluation. In *ICDM*, pages 521–528, 2001.

[46] D. Tikk, J. D. Yang, and S. L. Bang. Hierarchical text categorization using fuzzy relational thesaurus, Apr. 22 0.

[47] J. J. Verbeek. Supervised feature extraction for text categorization, Feb. 14 2002.

[48] 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.

[49] O. R. Zaïane and M.-L. Antonie. Classifying text documents by associating terms with text categories. In *Australasian Database Conference*, 2002.