Agrupación de documentos utilizando representaciones holográficas reducidas

Norma L. Cuautle-Rivera[1], Maya Carrillo[1] y Aurelio López-López[2]

[1] Facultad de Ciencias de la Computación, BUAP,
Av. San Claudio y 14 Sur Ciudad Universitaria, 72570 Puebla, México
n_lucero_c@hotmail.com, cmaya@cs.buap.mx
[2] Coordinación de Ciencias Computacionales, INAOE,
Luis Enrique Erro 1, Sta.Ma. Tonantzintla, 72840, Puebla, México
allopez@inaoep.mx

Resumen En este artículo proponemos la utilización de la representación holográfica reducida (HRR) para agrupar documentos de texto. Las HRRs son una representación novedosa de documentos que capturan información sintáctica de los mismos, la cual es producida utilizando la metodología del espacio vectorial conocida como Indización Aleatoria. Se empleó el conjunto de datos Reuters y se compararon los resultados con trabajos reportados en la bibliografía. Los resultados obtenidos muestran que las HRRs mejoran la tarea de agrupamiento con respecto a la representación vectorial empleando el mismo algoritmo de agrupamiento y tienen un desempeño competitivo con respecto a otros métodos de agrupamiento reportados, todo esto con una eficiencia en tiempo y espacio notablemente menor.

Palabras clave: Agrupamiento de documentos, modelo vectorial, indización aleatoria, representaciones holográficas reducidas.

1. Introducción

El ser humano emplea el lenguaje para expresar ideas y comunicarse con otros. La comprensión del lenguaje es compleja, debido a la variación y ambigüedad inherentes a él. Estas características dificultan el procesamiento automático del lenguaje, que se complica aún más porque la computadora no tiene aún la capacidad de comprensión del lenguaje que tenemos los humanos. En dicha comprensión interviene el contexto y la abstracción de ideas para desambiguar e interpretar de la mejor manera el significado de diferentes planteamientos. En las siguientes oraciones se ilustra la ambigüedad en distintos niveles del lenguaje:

a. A nivel léxico: Tomó una botella y se fue (¿bebió la botella o la tomó con la mano?).
b. A nivel morfológico: Nosotros plantamos papas (¿estamos en el proceso de plantar o ya se plantaron?).
c. A nivel sintáctico: Veo al gato con el telescopio (¿uso telescopio para ver al gato o veo al gato que tiene el telescopio?).

d. A nivel semántico: Todos los estudiantes de la escuela hablan dos lenguas (¿cada uno habla dos lenguas o sólo se hablan dos lenguas determinadas?).

El procesamiento del lenguaje natural (NLP) se apoya principalmente del análisis morfológico, léxico y sintáctico. Tradicionalmente, los documentos se representan como una lista de términos léxicos independientes, que son representados como vectores. Dichos vectores combinados linealmente permiten representar documentos. Esta representación se conoce como modelo de espacio vectorial (VSM), con el inconveniente de que en ella, se pierden las relaciones existentes entre palabras y, por lo tanto, la identificación de conceptos importantes que podrían ayudar a representar de manera más adecuada los documentos. Por otra parte cuando se utiliza el VSM los documentos y términos se representan en una matriz de **n x m**, donde **n** es el número de términos y **m** el número de documentos. Dado que no todos los términos aparecen en todos los documentos, esta matriz generalmente es muy dispersa y de dimensión grande, por lo que para procesarla es necesario emplear técnicas de reducción de dimensión. Uno de los métodos empleados es la descomposición en valores singulares (SVD) que es cara en términos de tiempo de procesamiento y memoria requerida. En el presente trabajo se experimenta con una técnica de reducción de espacio vectorial conocida como Indización Aleatoria (RI), que ha demostrado ser útil en tareas de recuperación de información y clasificación [5] [6] [7]. Con la ayuda de RI se construyen Representaciones Holográficas Reducidas (HRRs) [1] [2] [3] con el propósito de incluir información sintáctica en la representación de los documentos para determinar el efecto que dicha representación tiene en la tarea de agrupamiento de documentos. El resto del documento está organizado de la siguiente manera: en la sección 2 se explica la metodología de indización aleatoria; en la sección 3 se describe la representación holográfica reducida para documentos, en la 4 se mencionan algunos trabajos relacionados, en la sección 5 se describen los experimentos realizados; en la 6 se presentan los resultados obtenidos y finalmente la sección 7 presenta las conclusiones y posible trabajo futuro.

2. Indización aleatoria

La indización aleatoria (RI), es una metodología cuya idea básica es acumular vectores de contexto, basándose en la ocurrencia de las palabras en contextos. Esta técnica es incremental y no requiere una fase de reducción de dimensión [4]. La técnica de RI se describe en los siguientes pasos:

1. A cada contexto (documento) se le asigna una representación única y generada aleatoriamente, la cual es llamada vector índice. Estos vectores índice son dispersos, de alta dimensión, y ternarios, lo que significa que su dimensión (d) es del orden de miles, y que consisten de un pequeño número de +1s y -1s aleatoriamente distribuidos, con el resto de los elementos de los vectores iguales a 0.
2. Posteriormente se construyen vectores de contexto para cada palabra, mediante el recorrido del texto, cada vez que una palabra ocurre en un documento, el vector índice de dicho documento, se agrega al vector de contexto

para la palabra en cuestión. Las palabras son representadas por vectores de contexto d-dimensionales que son la suma de los contextos en los que aparecen dichas palabras.

3. Una vez construidos los vectores de contexto para el vocabulario (palabras diferentes en una colección de documentos), los documentos se representan como la suma ponderada de los vectores de contexto de las palabras que aparecen en ellos.

3. Representaciones holográficas reducidas

Las HRRs son vectores n-dimensionales, cuyos elementos siguen una distribución normal N(0,1/n). Dichos vectores permiten codificar estructura textual utilizando representaciones distribuidas. Para tal propósito se emplea el operador de convolución circular para relacionar términos, el cual es apropiado por las siguientes razones: al operar dos vectores se obtiene un vector del mismo tamaño, lo que permite su utilización en procesos recursivos sin incrementar la dimensión del espacio vectorial, además preserva la similitud estructural [1].

La convolución circular ($\otimes$) mapea dos vectores n-dimensionales a un vector $\mathbf{z}$. Si $\mathbf{x}$ y $\mathbf{y}$ son vectores n-dimensionales, entonces los elementos de $\mathbf{z} = \mathbf{x} \otimes \mathbf{y}$ son definidos como:

$$z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \qquad i = 0 \text{ to } n-1 \text{ (donde los subíndices son módulo-}n\text{)} \quad (1)$$

Para representar los documentos empleando HRRs se siguieron los siguientes pasos:

a. Se generaron vectores de contexto para el vocabulario de la colección utilizando la indización aleatoria.
b. Los documentos se etiquetaron por categorías sintácticas.
c. Se creó la representación HRR para cada término utilizando la convolución circular para relacionar el término con su categoría sintáctica. Por ejemplo, el par **deal/VBP** se codificó creandole un HRR $\mathbf{z}$ donde el vector $\mathbf{x}$ es el vector de contexto de deal y $\mathbf{y}$ un HRR generado aleatoriamente. Para cada categoría (etiqueta) sintáctica se generó un HRR aleatorio.
d. Los vectores (HRRs) resultantes se ponderaron empleando el esquema $tf - idf$ (que considera la importancia de los términos dentro de los documentos y de la colección) y se sumaron para representar documentos.

Si bien la HRR se ha utilizado para recuperar analogías [2], en clasificación [7] y recuperación de información [5], hasta donde se tiene conocimiento, no se ha utilizado para realizar agrupamiento de documentos.

4. Trabajo relacionado

Guan R. et al en [9] presentan un nuevo algoritmo de agrupamiento de texto llamado propagación de afinidad mediante semillas (en inglés Seeds Affinity

Propagation, SAP). Mencionan dos principales contribuciones, la primera es una nueva métrica de similitud, la cual captura información estructural de los textos; y la segunda es un método de construcción de semillas para mejorar el proceso de agrupamiento semi-supervisado. Utilizan el corpus Reuters y realizan una comparación de los resultados con los algoritmos k-Means y AP (Affinity Propagation). Obtienen mejores resultados con respecto a los algoritmos con que se comparan utilizando F-measure y la entropía como métricas de comparación. La información estructural que consideran mejora los resultados de agrupamiento empleando la medida de similitud que proponen, llamada Tri-conjunto (Tri-Set) en lugar del Coseno. Otro factor que contribuye a mejorar los resultados son las semillas seleccionadas, las cuales se obtienen de las características representativas de los objetos etiquetados.

Zimmerling M. en [10] emplea dos algoritmos llamados k-Means++ y KKz como extensiones del algoritmo k-Means básico, debido a que estos mejoran la elección de los centroides iniciales, mediante técnicas de semillas aleatorias. Para realizar las evaluaciones de los algoritmos utilizaron los conjuntos de datos Classic3, Reuters-21578 y F-measure como métrica de evaluación. Los algoritmos propuestos no mejoraron substancialmente al algoritmo k-Means en cuanto a tiempo de ejecución o número de iteraciones, debido a la naturaleza de los datos.

Cleuziou G., en [11] presenta un nuevo enfoque para determinar posibles cubiertas de datos para la agrupación de grupos traslapados. Propone un nuevo algoritmo, de nombre OKM, el cual es una generalización de k-Means. OKM consiste de una nueva función objetivo para minimizar bajo restricciones de multi-asignación, es decir, se explora el espacio de cubiertas en vez del espacio de particiones a diferencia de k-Means. Menciona que la tarea de asignar cada dato a uno o varios grupos no es una tarea trivial, por lo cual propone la heurística de desplazarse a través de la lista de grupos prototipo desde el más cercano al más lejano, y asignar cada vector mientras su imagen es mejorada, la nueva asignación es conservada si es mejor que la anterior, asegurándose de que disminuya la función objetivo. Los experimentos se realizaron con el corpus Reuters y se evaluaron con la métrica F-measure. Sus resultados muestran un comportamiento consistente del algoritmo OKM al proveer mejores grupos traslapados.

Su Z. et al, en [12], muestran un nuevo algoritmo de agrupamiento híbrido, el cual se basa en el algoritmo de cuantización vectorial (vector quantization, VQ) y el de estructura de crecimiento de celdas (growing -cell structure, GCS). Emplean VQ para mejorar la salida de agrupación de GCS y mejorar el problema de entrenamiento insuficiente. Los pesos de los nodos de salida del agrupamiento GSC son considerados como los vectores prototipo iníciales de VQ, después de varias ejecuciones de entrenamiento, los nuevos pesos de los vectores prototipo ganadores de VQ remplazarán a los pesos de los nodos de GCS. Tal proceso puede considerarse como una fase de entrenamiento adicional en la agrupación de GCS y resuelve el problema de entrenamiento insuficiente. Se denota como D el conjunto de documentos y F al conjunto de vectores de características de cada documento en D. Para cada nuevo documento insertado, no se requiere entrenar

la red desde el principio, debido a que las actualizaciones se pueden realizar basándose sobre resultados previos. Todos los documentos del corpus Reuters se ponderaron utilizando $tf-idf$, y se emplearon los algoritmos K-Means, VG, GCS y la propuesta hibrida. De acuerdo los resultados de los experimentos, el método propuesto alcanza mejor desempeño que los métodos con los que se comparan.

Ayad, et al. en [13], proponen combinar agrupaciones producidas por diferentes técnicas de agrupamiento para descubrir los tópicos de los documentos de texto, la agregación de estas agrupaciones genera mejores estructuras de datos. Después de que se forman los grupos de documentos, se emplea un proceso llamado extracción de tópicos, el cual selecciona los términos del espacio de características (es decir, el vocabulario de la colección entera) para describir el tópico de cada grupo, en esta etapa se re-calculan los pesos de los términos de acuerdo a la estructura de los grupos obtenidos. Para representar a los documentos se usó el modelo del espacio vectorial, para el pesado de los términos, se empleó $tf - idf$. Para la agrupación por agregación, se emplearon algoritmos jerárquicos, incrementales y particionales. Se utilizó F-measure para evaluar y comparar los tópicos extraídos y la calidad de la agrupación antes y después de la agregación. La evaluación experimental mostró que la agregación puede mejorar exitosamente tanto la calidad de la agrupación como la precisión de los tópicos comparándose con las técnicas de agrupación individuales.

5. Experimentos

Para evaluar el efecto de las HRRs en la tarea de agrupamiento, se utilizó el conjunto de documentos Reuters-21578 [8], el cual es popular en la comunidad de procesamiento de lenguaje natural, pues está constituido por noticias de diversos contextos que presentan diferentes características.

El corpus Reuters es una colección de 21,578 artículos que apareció en el servicio de noticias Reuters en 1987, distribuido sobre 116 categorías. Cada documento fue manualmente clasificado por los editores del servicio de noticias. El número de documentos asignado a cada categoría varía, algunas categorías tienen un gran número de documentos, como la llamada earn, mientras que otras categorías, como rye, tienen muy pocos documentos.

El primer experimento se realizó tomando el total de documentos de la colección con la misma cantidad de clases en los conjuntos de entrenamiento y prueba de Reuters, 68 clases. Posteriormente, con el fin de compararnos con los resultados reportados en [11], se utilizó un subconjunto de documentos distribuidos en 10 clases para contar con un total de 3696 documentos únicos, tratando de aproximarnos lo más posible al subconjunto de documentos utilizados en [11], hasta donde los detalles de dicha fuente nos permitió hacerlo, Tabla 1.

Dado que el corpus Reuters tiene clases muy desbalanceadas, se seleccionaron grupos de documentos para formar cinco conjuntos de documentos como se describe en [9]. Para lo cual se seleccionaron 800 documentos de texto, contenidos en 10 clases, para cada caso. La distribución de los diferentes números de documentos entre las 10 clases se muestra en la Tabla 2.

En Reuters los documentos de las clases *corn*, *grain* y *wheat* son muy similares entre sí, lo cual hace que sean muy difíciles diferenciarlos. La distribución de estas clases puede influir profundamente en los resultados del agrupamiento. Para cada uno de los cinco casos, las clases *corn*, *grain* y *wheat* contienen los siguientes porcentajes de documentos 10 %, 30 %, 50 %, 70 % y 90 %, respectivamente.

Se realizó el pre-procesado de cada documento del corpus etiquetando los términos con MontyLingua [14] para obtener las categorías sintácticas de los términos, se eliminaron los símbolos de puntuación y palabras vacías, y se truncaron las palabras utilizando el algoritmo de PorterStemmer.

Una vez preprocesados los documentos, se utilizó la indización aleatoria para generar las HRRs, empleando vectores de dimensión 1024.

Los experimentos fueron ejecutados en una PC Intel(R) Core(TM)2 Quad CPU 2.86 GHz con 4 GB de RAM. Se empleó la herramienta Weka 3.6.4 para realizar el agrupamiento de documentos y el algoritmo k-Means.

Tabla 1. Subconjuntos de documentos considerado en los primeros experimentos.

Conjunto de datos	*Documentos*	*Clases*
Total de documentos	9592	68
Subconjunto	3696	10

Tabla 2. Porcentajes de documentos por cada clase para los 5 casos considerados.

Casos	% *acq*	% *corn*	% *crude*	% *earn*	% *grain*	% *interest*	% *money-fx*	% *ship*	% *trade*	% *wheat*
Caso 1	30	2.5	20	15	5	10	7.5	5	2.5	2.5
Caso 2	10	10	10	10	10	10	10	10	10	10
Caso 3	8.8	16.3	7.5	6.3	16.3	5	3.8	2.5	16.3	17.5
Caso 4	7.5	23.1	6.3	5	23.1	3.8	1.3	2.5	3.8	23.8
Caso 5	2.5	30	1.3	1.3	30	1.3	1.3	1.3	1.3	30

6. Resultados

En la Tabla 3 se presentan los resultados obtenidos al comparar la propuesta de representación para los documentos, con los resultados obtenidos por diferentes métodos reportados en la bibliografía, hasta donde fue posible aproximar el número de clases y documentos. Puede verse que aparentemente los HRRs no mejorar los resultados del agrupamiento, sin embargo las grandes diferencias

que existen en las clases de la colección puede ser la principal causa. Tratando de homogeneizar el conjunto de pruebas se construyeron subconjuntos de documentos de acuerdo al trabajo de Guan R. et al [9]. En la Tabla 4 se presenta una comparación de los resultados obtenidos, al emplear las HRRs para representar documento y el algoritmo k-Means para clasificarlos, con los obtenidos en dicho trabajo. Puede observarse que la combinación de HRRs con k-Means se desempeñó mejor que el algoritmo AP cuando se utiliza el coseno como métrica de similitud, e incluso que el algoritmo AP con la métrica de similitud Tri-set propuesta en [9]. También mejoró al algoritmo SAP cuando se utiliza el coseno como métrica de similitud. Por otro lado, dicho algoritmo combinado con la métrica Tri-set mejora los resultados obtenidos con las HRRs. Mejora, que con mucha probabilidad se debe a la diferencia en las métricas de similitud empleadas, distancia euclidiana en k-Means.

El agrupamiento empleando HRRs, para representar los documentos y k-Means para agruparlos, mejora en cuanto a F-measure a k-Means cuando los documentos se representan con el VSM en un 48.63 % en promedio para los cinco casos considerados; en un 20. % a los obtenidos con AP y coseno; en un 27.70 % a los obtenidos con AP y Tri-Set y en un 7.72 % a SAP con coseno. Sin embargo es superado por SAP(Tri-set) en un 8.11 %.

Tabla 3. Resultados obtenidos por diferentes propuestas de agrupamiento

Referencia				*HRR*		
	Doc.	*Clases*	*F-measure*	*Doc.*	*Clases*	*F-measure*
1.An extended version of the k-Means method for overlapping clustering [11]	1308	10	0.76	3696	10	0.36
2.An Empirical Study of K-Means Initialization Methods for Document Clustering [10]	8193	65	0.35	9592	68	0.31
3.Topic discovery from text using aggregation of different clustering methods [13]	3000	10	0.39	3696	10	0.36

7. Conclusiones y trabajo futuro

Hemos presentado una representación novedosa para capturar la estructura sintáctica de documentos empleando la indización aleatoria. Los resultados mostraron que la representación es capaz de agrupar documentos de manera más precisa que cuando estos se representan con el VSM. Así mismo mostraron su competitividad con respecto a otros trabajos reportados en la bibliografía.

Tabla 4. Comparación de resultados con el trabajo relacionado, utilizando F-measure

Referencia	*Algoritmo*	*Caso 1*	*Caso 2*	*Caso 3*	*Caso 4*	*Caso 5*	*Prom.*
Text	K-MEANS	0.518	0.397	0.368	0.280	0.269	0.366
Clustering	AP(CC)	0.450	0.450	0.450	0.450	0.450	0.450
with Seeds	AP(Tri-Set)	0.577	0.482	0.419	0.364	0.290	0.426
Affinity	SAP (CC)	0.662	0.519	0.511	0.450	0.385	0.505
Propagation [9]	SAP(Tri-Set)	0.749	0.606	0.573	0.544	0.489	0.592
HRR-1024	K-MEANS	0.571	0.567	0.624	0.542	0.414	0.544

Utilizar RI y HRRs para representar documentos permite reducir la matriz de términos contra documentos en estos experimentos a únicamente una matriz de 800 por 1024 lo que optimiza el tiempo de cualquier algoritmo de agrupamiento. Para estos experimentos, el tiempo promedio fue de 0.26 min.

El trabajo en proceso nos permitirá probar la representación con otros algoritmos de agrupamiento y con diferentes métricas a fin de determinar de manera precisa el beneficio de la representación propuesta para el agrupamiento de documentos. También estamos por realizar un análisis cualitativo de la utilidad de la información sintáctica en la discriminación entre grupos, dado que aporta mayor información que el indexado tradicional en SVM.

Agradecimientos. Esta investigación fue financiada por el proyecto PROMEP /103.5/11/4481. El segundo y tercer autor parcialmente financiados por SNI México.

Referencias

1. Plate, T.A.: Holographic Reduced Representation: Distributed representation for cognitive structures. CSLI Publications, (2003)
2. Plate T.A : Analogy Retrieval and Processing with Distributed Vector Representation, Expert Systems, 17: 1, pp. 29-40,(2000)
3. Plate T. A. : Distributed Representation in: Encyclopedia of Cognitive Science, Macmillan Reference Ltd, (2002)
4. Sahlgren, M.: An Introduction to Random Indexing, In: Methods and Applications of Semantic Indexing Workshop at the 7th Int. Conf. on Terminology and Knowledge Engineering (2005)
5. Carrillo M., Eliasmith C., and Lopez-Lopez A., Combining Text Vector Representations for Information Retrieval, In: Text, Speech and Dialogue. Procs. of the 12th International Conference Text, Speech and Dialogue, LNAI, 5729 pp. 24-31,(2009)
6. Carrillo M., Villatoro-Tello E., López López A., Eliasmith C., Villaseñor-Pineda L., Montes-y-Gómez M.,Concept based representations for ranking in geographic information retrieval, In: Procs. of the 7th international conference on Advances in natural language processing, IceTal, pp. 85-96. (2010)

7. Sahlgren, M., Cöster R.: Using Bag-of-Concepts to Improve the Performance of Support Vector Machines in Text Categorization. In: Procs. of the 20th Int. Conf. on Computational Linguistics, pp. 487- 493 (2004)
8. Lewis D.D.: Reuters-21578 Text Categorization Test Collection. Available at http://www.daviddlewis.com/resources/testcollections/reuters 21578, May (2004)
9. Guan R., Shi X., Marchese M., Yang C., and Liang Y.: Text Clustering with Seeds Affinity Propagation. In: IEEE Trans. Knowledge and Data Eng., 23:4, pp.627-637, (2011)
10. Zimmerling M.: An Empirical Study of K-Means Initialization Methods for Document Clustering, Dresden University of Technology, Germany) at http://www.tik.ee.ethz.ch/ marcoz/other/clustering08.pdf, (2008)
11. Cleuziou G.: An extended version of the k-Means method for overlapping clustering. In: 19th ICPR Conference, Tampa, Florida, USA, pp. 1–4, (2008)
12. Su Z., Zhang L., Pan Y.: Document Clustering Based on Vector Quantization and Growing-Cell Structure, Developments. In: 16th international conference on industrial and engineering applications of artificial intelligence and expert systems, IEA/AIE 2003, Loughborough,UK, pp.326-336, (2003)
13. Ayad H., Kamel M.:Topic Discovery from Text Using Aggregation of Different Clustering Methods. In: Procs. of the 15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence Springer-Verlag London, UK, (2012)
14. Liu, Hugo: MontyLingua: An end-to-end natural language processor with common sense. web.media.mit.edu/∼hugo/montylingua (2004)