Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

Tipo de artículo: Artículo de revisión
Temática: Reconocimiento de patrones
Recibido: 05/10/2015 | Aceptado: 14/12/2015

Evaluación de diversas variantes de Indexado Aleatorio aplicadas a la categorización de documentos en el contexto del Aprendizaje en Línea

Preliminary assessment of Random Indexing variants for Text Categorization in Online Learning Context

Adrian Fonseca Bruzón[1*], Aurelio López López[2], José E. Medina Pagola[3]

[1]Centro de Estudios de Reconocimiento de Patrones y Minería de Datos. Datys. Ave. Patricio Lumumba s/n Altos de Quintero, Santiago de Cuba, Cuba
[2]Instituto Nacional de Óptica, Física y Electrónica. Sta María Tonantzintla, Puebla, México.
[3]Centro de Aplicaciones de Tecnologías de Avanzada. Datys. 7ma A #21406 e/ 214 y 216, Rpto. Siboney, Playa. La Habana, Cuba.

*Autor para correspondencia: adrian@cerpamid.co.cu

Resumen
El Indexado Aleatorio es una técnica de reducción de dimencionalidad que permite obtener un espacio de representación para las palabras a partir de un conjunto de contexto en los cuales éstas aparecen. Esta técnica es computacionalmente menos costosa en comparación con otras como LSI, PLSI o LDA. Estas características la convierten en una atractiva opción para ser empleada en ambientes de categorización de textos. En este trabajo comparamos varias variantes de Indexado Aleatorio al ser aplicadas a la tarea de categorización de textos. Los experimentos realizados en una subcolección del conjunto de datos Reuter-21578 muestran que el Indexado Aleatorio obtiene resultados alentadores, identificando algunas variantes que no muestran las ventajas necesarias para ser aplicadas en la tarea de interés.

Palabras claves: indexado aleatorio, categorización de textos, reducción de dimensionalidad

Abstract
Random Indexing is a recent technique for dimensionality reduction that allows to obtain a word space model from a set of contexts. This technique is less computationally expensive in comparison with others like LSI, PLSI or LDA. These characteristics turn it an attractive prospect to be used in text categorization. In this work, we compare several variants reported in the Random Indexing literature applied to text categorization task. Experiments conducted in a subcollection of the dataset Reuter-21578 show that Random Indexing produces promising results, identifying some versions without actual advantage for the task at hand.

Keywords: *random indexing, text categorization, dimensionality reduction*

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

Introducción

Hoy en día la personalización es un componente clave de muchos algoritmos de Aprendizaje en Línea o Sistemas de Recomendación. Usualmente estos algoritmos crean un perfil de usuario para representar las necesidades de información de los usuarios. Estos algoritmos tienen que decidir por cada documento cuándo se ajusta al perfil del usuario o no.

En la Minería de Textos, estos métodos son particularmente importantes si tomamos en consideración el enorme volumen de nueva información que cada día es generada en Internet. En estos algoritmos un componente fundamental es el algoritmo de clasificación empleado. Sin embargo, en esta tarea, estos algoritmos tienen que lidiar con dos grandes problemas, el lenguaje y la dimensionalidad. El lenguaje natural es un gran reto para las Ciencias de la Computación. Por un lado las palabras son ambiguas, es decir una palabra puede tener diversos significados y varias palabras pueden ser empleadas para referirse a un mismo concepto. Por otra parte, en el contexto del Aprendizaje en Línea, los documentos arriban continuamente, y usualmente ellos contienen nuevos términos no vistos que deben ser tenidos en cuenta para análisis posteriores.

El otro problema es la dimensionalidad del espacio de representación. Usualmente, los documentos son representados por medio de un vector de una dimensión igual al tamaño del vocabulario de la colección, o en un entorno real igual al número de palabras vistas hasta el momento. Esta situación afecta significativamente el desempeño de los algoritmos de Aprendizaje en Línea y de Categorización de Textos de forma general.

Algunos algoritmos han sido reportados en la literatura con el objetivo de resolver uno o varios de los problemas anteriormente expuestos. Entre ellos podemos encontrar el Indexado de Semántica Latente (LSI) (DUMAIS et al., 1995), el Indexado Probabilístico de Semántica Latente (PLSI) (HOFMANN, 1999), o la Asignación Latente de Dirichlet (LDA) (BLEI et al., 2003). Sin embargo estos métodos son computacionalmente costosos, o requieren de cargar completamente en memoria la matriz de frecuencias términos-documentos. Estas limitaciones reducen su aplicabilidad en ambientes de Aprendizaje en Línea donde ocurren actualizaciones frecuentes en la información disponible.

El Indexado Aleatorio (SAHLGREN, 2005) puede constituir una alternativa viable, dado que este método es computacionalmente menos costoso y no requiere del acceso en memoria de toda la matriz de frecuencias términos-documentos. Por estas razones, este método es más atractivo para ser empleado en un ambiente en línea. Por otra parte, varias variantes diferentes del Indexado Aleatorio han sido reportadas en la literatura con el objetivo de resolver diversas tareas del Procesamiento del Lenguaje Natural (PLN).

En este trabajo presentamos una comparación experimental de varias de estas variantes en el contexto de la categorización de documentos para la tarea de Aprendizaje en Línea. Los resultados obtenidos indican que esta representación puede producir resultados competitivos con vectores de una baja dimensión.

El resto de este artículo está organizado de la forma siguiente: en la sección siguiente se describe el Indexado Aleatorio y sus variantes fundamentales. Luego se presenta nuestra propuesta de emplear el Indexado Aleatorio en el contexto del Aprendizaje en Línea. Seguidamente describimos el marco experimental y discutimos los resultados obtenidos. Finalmente, proveemos nuestras conclusiones y posibles áreas para el trabajo futuro.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

Indexado Aleatorio

El Indexado Aleatorio (SAHLGREN, 2005) fue introducido por Pentti Kanerva *et al* en el 2000 (KANERVA et al., 2000) y está basado en tres presupuestos fundamentales:

- Hipótesis de distribución: Palabras con significados similares aparecen en contextos similares (RUBENSTEIN and GOODENOUGH, 1965).
- Lema Johnson-Lindenstrass: La proyección de un espacio de alta dimensionalidad en un espacio de una dimensión mucho menor puede ser realizada de forma tal que la distancia entre los puntos del espacio sea prácticamente preservada (JOHNSON and LINDENSTRAUSS, 1984).
- Existen muchas más direcciones pseudo-ortogonales que direcciones realmente ortogonales en un espacio de una alta dimesionalidad (HECHT-NIELSEN, 1994).

Las ideas de Kanerva fueron desarrolladas por Magnus Sahlgren del Instituto Sueco de Ciencias de la Computación. Él formalizó el Indexado Aleatorio como un proceso de dos pasos de la siguente forma:

1. Primeramente, a cada contexto (por ejemplo un documento o una palabra) le es asignado una representación única generada de forma aleatoria llamada *vector índice*. Estos vectores índices son dispersos, de una dimensión alta, y ternarios, lo que significa que su dimensión (d) se encuentra en el orden de los miles, y que están compuestos por un número pequeño de +1 y -1 distribuidos aleatoriamente, con el resto de los elementos del vector puestos en 0.
2. Luego, los *vectores de contexto* son construidos escaneando a través del texto, y cada vez que aparece una palabra en el contexto (por ejemplo en un documento, o dentro de una ventana deslizante), el vector índice d-dimensional del contexto es adicionado al vector de contexto de la palabra de interés. De esta forma las palabras son representadas por un vector de contexto d-dimensional que es construido como la suma de las palabras que forman el contexto en el que aparece la palabra en cuestión.

Diferentes tipos de contexto pueden ser empleados durante el proceso de construcción del Indexado Aleatorio. Los más ampliamente empleados son considerar todo el documento como contexto o tomar términos como contexto. Cuando los términos son considerados como contexto, usualmente se emplea una ventana alrededor del término que se está analizando. En este último caso, el vector de contexto en actualizado con los vectores índice de aquellos términos que se encuentran en la vecindad del término objetivo.

Otra forma de emplear los términos como contexto fue presentada en (MUSTO, 2010). En este trabajo, un vector índice es asignado a cada término. En este caso, el vector de contexto es actualizado con todos los vectores índices de los términos que se encuentran en el documento.

El Indexado Aleatorio captura la semántica de los términos basado en las coocurrencias. Sin embargo, Cohen *et. al.* concluyen que esta técnica presenta algunos inconvenientes para determinar relaciones indirectas entre las palabras (COHEN et al., 2010). Para superar esta limitación, ellos proponen una extensión nombrada Indexado Aleatorio Reflexivo. En esta extensión, ellos asignan un vector índice a cada término, luego se

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

obtiene la representación del documento d como la suma de los vectores índices de aquellos términos que aparecen en d. En lo sucesivo, estos vectores de los documentos son empleados para construir los vectores de contexto de los términos. Este proceso puede ser repetido varias veces, pero de acuerdo a sus experimentos, los mejores resultados se obtienen luego de una o dos iteraciones.

Una vez construidos los vectores de contexto, podemos obtener la representación para un documento d, adicionando los vectores de contexto de aquellos términos que aparecen en él (SAHLGREN and CÖSTER, 2004). Durante este proceso, los vectores de contexto pueden ser multiplicados por el peso que indica la importancia relativa de cada término en el documento.

Durante el proceso de obtener la representación final para los documentos, podemos realizar algunas transformaciones sobre los vectores de contexto; en particular Higging y Burstain proponen restar la media de los vectores de contexto a cada vector de contexto antes de obtener la representación de los documentos (HIGGINS and BURSTEIN, 2007). De acuerdo a los autores, en el Indexado Aleatorio la semejanza entre los documentos tiende a incrementarse conforme crece su longitud, independientemente de su relación. Con esta transformación, ellos intentan mitigar este inconveniente.

Aprendizaje en Línea con Indexado Aleatorio

La mayoría de los trabajos reportados en la literatura han empleado el tradicional modelo de espacio vectorial, también conocido como bolsa de palabras, para representar a los documentos. Sin embargo, es conocido que este modelo no puede capturar las relaciones semánticas que existen entre los términos que forman un documento.

Por otra parte, varias tareas como el Filtrado de Información, la Recomendación de Noticias y la Categorización de documentos se ven beneficiadas por el empleo de técnicas que no asumen que los términos presentes en un documento son independientes entre sí.

En este escenario, el Indexado Aleatorio es una representación plausible, teniendo la ventaja de ser menos costosa computacionalmente que otras técnicas como LSI, PLSI o LDA.

En este estudio exploramos diferentes variantes reportadas en la literatura relacionada con el Indexado Aleatorio en el contexto de la categorización de textos como parte del Aprendizaje en Línea. Estos modelos emplean usualmente un perfil de usuario en su modelación. Un perfil es la representación interna de las necesidades de información de un usuario. Generalmente, esta tarea es modelada como un proceso de clasificación binaria donde el clasificador debe decidir por cada nuevo documento cuándo este es similar o no al perfil del usuario.

Una representación simple de un perfil de usuario es creada mediante la suma en un único vector de todos aquellos documentos que son relevantes para el usuario. Siguiendo esta misma idea, podemos igualmente construir un vector para representar la información irrelevante.

Con esta representación, cada nuevo documento es clasificado como relevante para el usuario si la semejanza con el vector que representa a los documentos relevantes es superior con respecto a la semejanza al vector que representa a los documentos irrelevantes.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

Cuando se emplea un modelo semántico como el Indexado Aleatorio, se requiere de un paso extra. Durante la etapa de entrenamiento, es necesario considerar en el perfil toda la información disponible en la colección de entrenamiento para la construcción del modelo semántico. Este modelo semántico será empleado para representar tanto los documentos de entrenamiento como aquellos nuevos documentos a ser clasificados.

Experimentos

Para la experimentación empleamos la colección de documentos Reuter-21578[1]. Varios subconjuntos han sido creados a partir de esta colección, de ellos los más conocidos son:

- El conjunto de las 10 categorías con el mayor número de muestras de entrenamiento.
- El conjunto de las 90 categorías con al menos una muestra de entrenamiento y una en el conjunto de pruebas.
- El conjunto de las 115 categorías con al menos una muestra en el conjunto de entrenamiento.

En particular para este estudio seleccionamos el primero de estos subconjuntos. En la Tabla 1 se muestra el número de muestras de entrenamiento para cada clase.

Tabla 1. Número de documentos de entrenamiento por clases.

Clases	**Número de documentos de entrenamiento**
earn	3753
acq	2131
wheat	264
money-fx	600
corn	206
trade	449
grain	527
interest	389
crude	510
ship	276

Al inspeccionar la tabla podemos notar que existe una acentuada diferencia en el número de muestra de cada una de las clases. En particular las clases earn y acq contienen un número bastante mayor de muestras en comparación con el resto de las clases. Dado que el objetivo de este estudio noes analizar los efectos que pueden

[1]http://www.daviddlewis.com/resources/testcollections/reuters21578/

ser introducidos por el desbalance entre las clases, en nuestros experimentos decidimos ignorar las clases earn y acq, considerando finalmente las 8 clases restantes.

Durante el preprocesamiento de los documentos, las etiquetas y las palabras de parada fueron eliminadas y fue aplicado un proceso de lematización. Finalmente fue empleado el esquema de pesado de términos TF-IDF.

Los experimentos fueron realizados con el fin de comparar el desempeño obtenido con las diversas variantes de Indexado Aleatorio. Con este fin, desarrollamos un esquema de validación cruzada con 5 particiones.

Para cada clase, el perfil del usuario es construído por medio de dos vectores. Uno de ellos para representar aquellos documentos que son relevantes para el usuario y el otro para los que no son de su interés. Cada uno de estos vectores fue construído adicionando todos los vectores que pertenecen, o no, a la clase en el conjunto de entrenamiento.

Durante el proceso de clasificación, un documento es etiquedo como Relevante para un perfil si su similaridad con respecto al vector que representa a los documentos relevantes es superior a la obtenida con respecto al vector que representa a los documentos no relevantes.

Como medida de evaluación seleccionamos la tradicional medida *precisión*, es decir la proporción de documentos clasificados como Relevantes que realmente son relevantes, así como la medida *relevancia*, es decir la proporción de documentos relevantes que realmente son clasificados como Relevantes. Estas medidas son usualmente combinadas en la popular medida F_1.

$$F_1 = \frac{2 * precisión * relevancia}{precisión + relevancia}$$

Dado que la medida F_1 se calcula de forma separada para cada clase, consideramos como medida de evaluación global la media obtenida sobre todas las clases, comúnmente conocida como $Macro - F_1$.

En la Figura 1 y la Tabla 2 se muestra la media de los resultados obtenidos en cada una de las corridas con la medida $Macro - F_1$.

En los resultados, RI representa el empleo del Indexado Aleatorio cuando son considerados los documentos como contexto; de la misma forma, wRI cuando son considerados los términos como contexto y se emplea una ventana alrededor del término objetivo y TRI cuando no es empleada ventana alguna. Por último, RRI se refiere al empleo del Indexado Aleatorio Reflexivo. Aquellos modelos que presentan el sufijo "-MV" representan a los resultados obtenidos cuando la media de los vectores de contexto es restada de los vectores de contexto antes de construir la representación del documento.

Para el Indexado Aleatorio consideramos vectores de tamaño 5000, con 5 posiciones seleccionadas como +1 y 5 posiciones seleccionadas como -1, cuando son generados los vectores índices. Para el modelo wRI consideramos una ventana de tamaño 2 alrededor de la palabra. En el caso del modelo RRI solamente se realizó una iteración.

Revista Cubana de Ciencias Informáticas
Vol. 10, No. 1, Enero-Marzo, 2016
ISSN: 2227-1899 | RNPS: 2301
http://rcci.uci.cu
Pág. 162-171

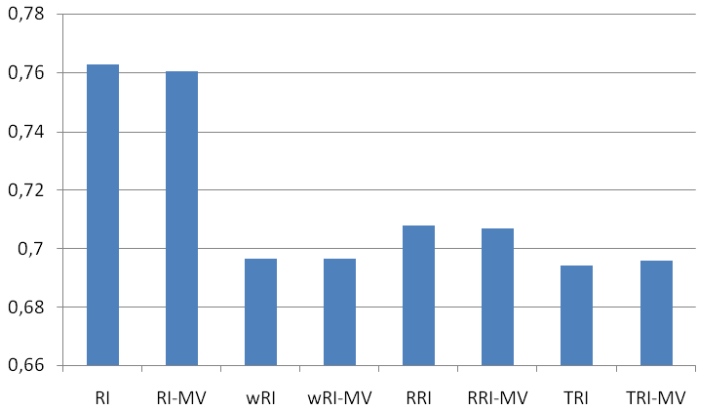

Figura 1. Promedio de la medida $Macro - F_1$.

Tabla 2. Promedio de la medida $Macro - F_1$ para varias variantes de Indexado Aleatorio.

Variante	**Promedio de la medida** $Macro-F_1$
RI	0.7626
RI-MV	0.7604
wRI	0.6964
wRI-MV	0.6964
RRI	0.7078
RRI-MV	0.7068
TRI	0.6942
TRI-MV	0.6958

Discusión

De los resultados podemos notar varios comportamientos. Primeramente, cuando las diversas variantes de Indexado Aleatorio son comparadas, los mejores resultados son obtenidos cuando los documentos son considerados como contexto. En este caso, esta variante es superior con respecto al resto de las variantes analizadas en aproximadamente en un 8 % - 9 %.

Otro aspecto relevante es que los resultados alcanzados con el Indexado Aleatorio Reflexivo son superiores a los resultados obtenidos con el Indexado Aleatorio cuando los términos son considerados como contexto. El Indexado Aleatorio Reflexivo fue propuesto para capturar las relaciones indirectas entre los términos; sin embargo, su utilidad no es la misma en otras tareas como es el caso de la analizada en este trabajo.

Podemos observar además que, en la mayoría de los casos, cuando se substrae la media de los vectores de

contexto a los vectores de contexto antes de obtener la representación final de los documentos no se obtiene una ganancia consistente que verdaderamente justifique su empleo. Por tal motivo, no encontramos una razón que justifique la incorporación de esta operación dado su aporte actual.

Finalmente, la ventaja fundamental del Indexado Aleatorio es que solamente consideramos vectores de 5000 elementos. Este aspecto toma particular relevancia si tomamos en consideración el objetivo de aplicar el Indexado Aleatorio en la tarea del Aprendizaje en Línea, donde cada nuevo documento puede contener nuevos términos no vistos con anterioridad. Con el Indexado Aleatorio, el problema de que frecuentemente aparezcan nuevos términos no afecta la eficiencia dado que los documentos son siempre representados con vectores de una dimensión fija.

Conclusiones

El Indexado Aleatorio es una técnica de indexado que de forma implícita realiza un proceso de reducción de la dimencionalidad, y en su sencillo proceso iterativo puede adquirir las relaciones semánticas que existen entre los términos. Varios enfoques han sido reportados en la literatura para el Indexado Aleatorio, aplicados a diversas tareas del Procesamiento del Lenguaje Natural y la Minería de Textos. Es este trabajo reportamos la comparación de las variantes más relevantes del Indexado Aleatorio aplicadas a la tarea del Aprendizaje en Línea. Los resultados reportados muestran que considerar los documentos como contextos permiten obtener los mejores resultados, aún con vectores de aproximadamente un tercio de la cantidad de términos total del conjunto de entrenamiento. Sin embargo, aún queda por analizar el impacto del tamaño del conjunto de entrenamiento; considerando que en un modelo semántico, varios documentos son necesarios para poder obtener una representación válida de las relaciones que existen entre los términos.

En esta dirección, en trabajos futuros analizaremos el comportamiento del Indexado Aleatorio en relación con la cantidad de datos disponibles para su construcción. Además, planeamos evaluar cómo influye el problema del desbalance entre las clases en el comportamiento del Indexado Aleatorio.

Referencias

BLEI, D. M., NG, A. Y., and JORDAN, M. I. (2003). Latent dirichlet allocation. *the Journal of machine Learning research*, 3:993–1022.

COHEN, T., SCHVANEVELDT, R., and WIDDOWS, D. (2010). Reflective random indexing and indirect inference: A scalable method for discovery of implicit connections. *Journal of Biomedical Informatics*, 43(2):240 – 256.

DUMAIS, S., FURNAS, G., LANDAUER, T., DEERWESTER, S., DEERWESTER, S., and OTHERS (1995). Latent semantic indexing. In *Proceedings of the Text Retrieval Conference*.

HECHT-NIELSEN, R. (1994). Context vectors: general purpose approximate meaning representations self-organized from raw data. *Computational intelligence: Imitating life*, pages 43–56.

HIGGINS, D. and BURSTEIN, J. (2007). Sentence similarity measures for essay coherence. In *Proceedings of the 7th International Workshop on Computational Semantics*, pages 1–12.

HOFMANN, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM.

JOHNSON, W. B. and LINDENSTRAUSS, J. (1984). Extensions of lipschitz mappings into a hilbert space. *Contemporary mathematics*, 26(189-206):1.

KANERVA, P., KRISTOFERSSON, J., and HOLST, A. (2000). Random indexing of text samples for latent semantic analysis. In *Proceedings of the 22nd annual conference of the cognitive science society*, volume 1036.

MUSTO, C. (2010). Enhanced vector space models for content-based recommender systems. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys '10, pages 361–364. ACM.

RUBENSTEIN, H. and GOODENOUGH, J. B. (1965). Contextual correlates of synonymy. *Communications of the ACM*, 8(10):627–633.

SAHLGREN, M. (2005). An introduction to random indexing. In *Proceedings of the Methods and Applications of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, TKE 2005*.

SAHLGREN, M. and CÖSTER, R. (2004). Using bag-of-concepts to improve the performance of support vector machines in text categorization. In *Proceedings of the 20th international conference on Computational Linguistics*, page 487. Association for Computational Linguistics.