Agrupación de Documentos Utilizando Representaciones Holográficas Reducidas

por

Norma Lucero Cuautle Rivera

Tesis profesional para obtener el título de:

Licenciado en Ciencias de la Computación

en la

Benemérita Universidad Autónoma de Puebla
Facultad de Ciencias de la Computación

Junio, 2012
Puebla, Pue.

Asesor: Dra. Maya Carrillo Ruíz
Co-asesor: Dr. Aurelio López López

Resumen

En la presente tesis planteamos el uso de representaciones holográficas reducidas (HRRs) en la tarea de agrupamiento de 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. Para la evaluación de la representación propuesta 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.

Dedico esta tesis a mi mamá Catalina, quien se merece todo mi respeto y admiración, gracias a su apoyo tuve la perseverancia de terminar este trabajo. A mis hermanas Adriana y Marisol, por su comprensión y sustento. A mi novio Carlos, por su apoyo incondicional, valiosos consejos y por ser la inspiración en mi vida.

Agradezco de manera especial a mi asesora, por el tiempo dedicado a este trabajo, por la paciencia que me tuvo a lo largo de la realización de esta tesis y por los consejos e ideas acertadas. También expreso mi sincero agradecimiento a mi co-asesor por sus valiosas sugerencias y tiempo empleado en esta tesis.

Índice general

Lista de figuras **9**

Lista de tablas **11**

1. Introducción **13**
1.1. Problemática 13
1.2. Hipótesis 14
1.3. Objetivos generales y específicos 14
1.3.1. General. 14
1.3.2. Particulares. 14

2. Conceptos básicos **17**
2.1. Agrupamiento 17
2.1.1. Tipos de métodos de agrupamiento 19
2.2. Modelo de espacio vectorial 23
2.3. Indización aleatoria 24
2.4. Representaciones holográficas reducidas 25

3. Trabajo Relacionado **27**

4. Metodología **31**

5. Experimentos **35**
5.1. Corpus 35
5.2. Métricas de evaluación 36
5.3. Entorno experimental 37

6. Resultados **41**

7. Conclusiones **45**
7.1. Principales resultados 45
7.2. Trabajo futuro 45

A. Publicación derivada de la tesis **47**

Índice de figuras

2.1. Ejemplo de la formación de tres grupos. 19
2.2. Una taxonomía de los enfoques de agrupamiento en [8]. 19
2.3. El dendograma obtenido al usar un método aglomerativo en [8]. 20

4.1. Diagrama a bloques del modelo propuesto 31

5.1. Porcentajes de documentos de cada una de las 10 clases seleccionadas como en [12] 38

Índice de tablas

3.1. Trabajos relacionados. 30

4.1. Lista de categorías sintácticas. 32

5.1. Subconjunto de documentos considerados en los primeros experimentos. 38
5.2. Porcentajes de documentos por cada clase entre los 5 casos descritos en [12] 39

6.1. Resultados de agrupamiento con dimensiones de 1024 y 2048 componentes 41
6.2. Resultados de tres algoritmos de agrupamiento 42
6.3. Resultados obtenidos de las diferentes propuestas de agrupamiento 43
6.4. Comparación de resultados con el trabajo relacionado, utilizando F-measure 43
6.5. Tiempos de arupamiento con HRR y K-Means 44

Capítulo 1

Introducción

1.1. Problemática

El ser humano emplea el lenguaje natural 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 natural, lo cual se complica aún más debido a que la computadora no tiene la capacidad de comprensión propia del ser humano. Con dicha comprensión utilizamos el contexto y la abstracción de ideas para desambiguar e interpretar de mejor manera el significado de ciertos planteamientos. En las siguientes oraciones se ilustra la ambigüedad en distintos niveles del lenguaje:

1. A nivel léxico: Tomó una botella y se fue (¿bebió la botella o la tomó con la mano?).

2. A nivel morfológico: Nosotros plantamos papas (¿estamos en el proceso de plantar o ya se plantaron?).

3. A nivel sintáctico: Veo al gato con el telescopio (¿uso telescopio para ver al gato o veo al gato que tiene el telescopio?).

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

Los principales avances en el área del procesamiento del lenguaje natural(NLP) se han apoyado en el procesamiento de los niveles bajos del lenguaje

i.e. análisis morfológico, léxico y sintáctico, con poco entendimiento del nivel semántico. Tradicionalmente, los documentos se constituyen como una lista de términos léxicos independientes que son representados como vectores. Posteriormente, dichos vectores, mediante combinaciones lineales, permiten representar documentos. Este modelo se conoce como modelo de espacio vectorial (VSM). Sin embargo, en dicha representación se pierde cualquier relación existente entre palabras y por lo tanto la identificación de conceptos importantes para representar documentos, además se obtienen vectores de dimensión considerablemente grande, lo cual genera problemas computacionales para el procesamiento de los mismos. Por otra parte, también se han explorado métodos que reducen la dimensión del espacio generado inicialmente por el VSM. En la presente investigación se busca estudiar el efecto que la representación holográfica reducida (HRR) [15, 13, 14] tiene en la efectividad de la tarea de agrupamiento. Se pretende utilizar diferentes algoritmos de agrupamiento. Se emplearán a las HRRs para capturar información sintáctica de los documentos. Adicionalmente, se experimentará con una técnica de reducción de dimensión, conocida como indización aleatoria (RI), que ha demostrado ser útil en tareas de recuperación de información y clasificación [8, 3].

1.2. Hipótesis

Representaciones holográficas reducidas son útiles en la tarea de agrupamiento de documentos.

1.3. Objetivos generales y específicos

1.3.1. General.

Establecer si la representación holográfica reducida es útil en la tarea de agrupamiento cuando se incluye información sintáctica de los documentos, y de esta forma enriquecer su representación con respecto al tradicional modelo del espacio vectorial.

1.3.2. Particulares.

- Evaluar la HRR en agrupamiento.

- Explorar si la indización aleatoria utilizada para reducción de la dimensión vectorial es útil en la tarea de agrupamiento.
- Determinar la utilidad de la HRR empleando al menos tres algoritmos de agrupamiento.

Capítulo 2

Conceptos básicos

En este capítulo se describe el concepto de agrupamiento de documentos, así como su taxonomía y los tipos de métodos de agrupamiento, con el fin de comprender el objetivo de dicha tarea y los algoritmos empleados en esta tesis. Se explicara el modelo de espacio vectorial, la indización aleatoria, utilizada para la reducción de la dimensión vectorial, y por último presenta la representación holográfica reducida propuesta en esta investigación para representar documentos y explorar su utilidad en la tarea de agrupamiento.

2.1. Agrupamiento

El agrupamiento (clustering) es considerado como un problema de aprendizaje no supervisado, debido a que se trata de encontrar una estructura en una colección de datos no etiquetados. El proceso de agrupamiento consiste en organizar objetos en grupos (clusters) cuyos miembros son similares de alguna manera. Intuitivamente los objetos del mismo grupo son similares entre sí y diferentes a los objetos de otros grupos. En la Figura 2.1 vemos un ejemplo de cómo un conjunto de datos son agrupados en 3 grupos distintos.

A continuación describimos algunos de los requerimientos que se deben satisfacer al realizar agrupamiento:

- Escalabilidad.
- Capacidad de trabajar con diferentes tipos de atributos.
- Descubrimiento de grupos con forma arbitraria.

- Requerimientos mínimos para el conocimiento del dominio para determinar los parámetros de entrada.
- Capacidad para tratar con el ruido y los valores atípicos.
- Insensibilidad con respecto al orden de los registros de entrada.
- Dimensionalidad alta
- Interpretabilidad y usabilidad.

Existen varios problemas con agrupamiento, tales como:

- Las técnicas actuales de agrupamiento no se ocupan de todos los requerimientos de forma adecuada (y simultáneamente).
- Hacer frente a un gran número de dimensiones y el gran número de elementos de datos puede ser problemático debido a la complejidad en tiempo.
- La eficiencia del método depende de la definición de "distancia" (basada en la distancia de agrupamiento).
- Si no existe una medida de distancia se debe definir, lo cual no siempre es fácil, especialmente en espacios multidimensionales.
- El resultado del algoritmo de agrupamiento "que en muchos casos puede ser arbitrario en sí" puede ser interpretado de diferentes maneras.

El agrupamiento tiene varias aplicaciones, tales como recuperación de documentos, aplicaciones de minería de datos y base de datos espaciales (como sistemas de información geográfica (GIS) o datos procedentes de astronomía), *marketing*, diagnóstico médico, análisis de ADN en biología computacional, entre otras.

Jain y Dubes en [8] mencionan en la taxonomía de los algoritmos de agrupamiento una distinción entre enfoques jerárquicos y particionales, ver Figura 2.2. Los primeros producen una serie anidada de particiones cuya estructura puede ser representada por medio de un árbol, mientras los segundos tratan de particionar el conjunto de datos en grupos. Estos enfoques se definen en las siguientes secciones.

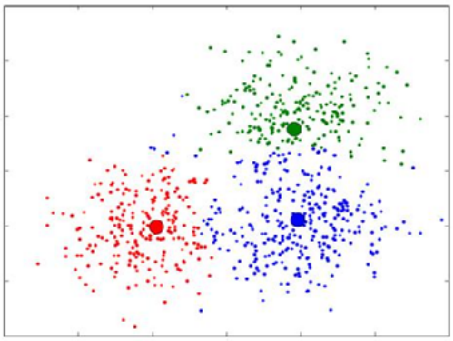

Figura 2.1: Ejemplo de la formación de tres grupos.

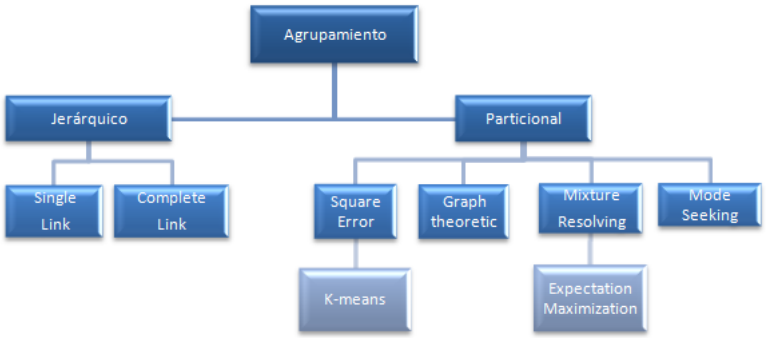

Figura 2.2: Una taxonomía de los enfoques de agrupamiento en [8].

2.1.1. Tipos de métodos de agrupamiento

Existen diferentes métodos para el agrupamiento de documentos, la elección de uno de ellos, depende tanto del tipo de agrupos deseados, así como el desempeño del método de acuerdo al tipo de datos con el que se trabaja. A continuación definimos algunos métodos de agrupamiento de datos de acuerdo a [8].

Métodos jerárquicos

Los métodos jerarquicos puden dividirse en aglomerativos y divisivos, los cuales se describen a continuación:

Aglomerativos: inician con cada elemento en su propio grupo y encuentran los pares más similares para unirlos en un nuevo grupo, sucesivamente hasta que los grupos forman un solo grupo.

Divisivos: inician con todos los elementos en un solo grupo y recursivamente dividen cada grupo en grupos más pequeños.

Un algoritmo jerárquico produce un dendograma, el cual representa el agrupamiento anidado de objetos y los niveles de similitud en los que los grupos cambian. Un ejemplo de dendograma se muestra en la Figura 2.3. El dendograma puede ser dividido en diferentes niveles para producir diferentes grupos de los datos.

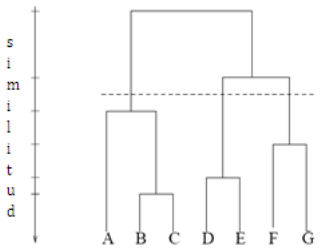

Figura 2.3: El dendograma obtenido al usar un método aglomerativo en [8].

Los algoritmos jerárquicos más populares son single-link y complete-link, los cuales difieren en la manera en que caracterizan la similitud entre un par de grupos. En single-link la distancia entre dos grupos es la mínima de las distancias entre todos los pares de objetos dibujados de los dos grupos. En el algoritmo complete-link la distancia entre dos grupos es la máxima de las distancias de todos los pares entre los objetos en los dos grupos. Desde el punto de vista práctico, se ha observado que el algoritmo complete-link produce jerarquías más útiles en muchas aplicaciones que el algoritmo single-link.

A continuación se explican los pasos que siguen los algoritmos single link y complete-link.

Single-link

1. Colocar cada objeto en su propio grupo. Construir una lista de distancias para todos los distintos pares de objetos desordenados, y ordenar

esta lista en orden ascendente.

2. Recorrer la lista ordenada de distancias, formando un grafo para cada distinto valor de disimilitud d_k en donde los pares de objetos más cercanos a d_k están conexos por una arista del grafo. Si todos los objetos son miembros de un grafo conexo, detenerse. De otra forma, repetir este paso.

3. La salida del algoritmo es una jerarquía anidada de grafos, los cuales pueden ser conexos en algún nivel de disimilitud formando una partición (grupo) identificado por componentes conexos en el grafo correspondiente.

Complete-link

1. Colocar cada objeto en su propio grupo. Construir una lista de distancias para todos los distintos pares de objetos desordenados, y ordenar esta lista en orden ascendente.

2. Recorrer la lista ordenada de distancias, formando un grafo para cada distinto valor de disimilitud d_k en donde los pares de objetos más cercanos a d_k están conexos por una arista del grafo. Si todos los objetos son miembros de un grafo completamente conexo, detenerse.

3. La salida del algoritmo es una jerarquía anidada de grafos, los cuales pueden ser cortados en algún nivel de disimilitud formando una partición (grupo) identificado por componentes conexos en el grafo correspondiente.

Métodos particionales

Los métodos particionales construyen k particiones a partir de un conjunto de datos T, donde cada partición representa un grupo y los k grupos deben satisfacer:

i) Cada grupo contiene al menos un objeto.

ii) Cada objeto debe pertenecer a un solo grupo.

Los métodos particionales tienen ventajas en aplicaciones que involucran grandes conjuntos de datos, evitando la construcción de un dendograma, lo cual, es computacionalmente costoso para grandes conjuntos de datos. Uno de los problemas en el uso de algoritmos particionales es la elección del número deseado de grupos de salida. Dubes en [2] provee una guía sobre esta importante decisión.

Entre los algoritmos particionales tenemos a k-Means, EM y Farthest-First. A continuación describimos los paso que siguen estos algoritmos:

a) K-Means

   a) Elegir k centros de grupos para coincidir con k puntos aleatoriamente definidos dentro del conjunto de objetos.

   b) Asignar cada objeto al centro del grupo más cercano.

   c) Recalcular los centros de los grupos usando la actual membresía.

   d) Si no es conocido un criterio de convergencia, ir al paso *a*). Algunos criterios de convergencia típicos pueden ser: no reasignar los objetos a los nuevos centros de los grupos, o el descenso mínimo del error cuadrado.

   K-Means es un algoritmo para agrupar objetos en función de sus atributos o características en k número de grupo. La agrupación se realiza minimizando la suma de los cuadrados de las distancias entre los datos y el centroide del correspondiente grupo.

   La función más intuitiva, y frecuentemente usada, en las técnicas de agrupamiento particional es el error cuadrado, la cual tiende a trabajar bien con los grupos aislados y compactos. El error cuadrado para un grupo L de un conjunto de objetos H (conteniendo k grupos) es:

$$e^2(H, L) = \sum_{j=1}^{k} \sum_{i=1}^{n_j} ||x_i^{(j)} - c_j||^2 \tag{2.1}$$

Donde $x_i^{(j)}$ es el i-ésimo objeto perteneciente al j-ésimo grupo, c_j es el centroide del j-ésimo grupo y n_j es el total de objetos en el j-ésimo grupo.

b) Expectation Maximization

El algoritmo expectation maximization (EM) se basa en el modelo de las mezclas finitas. Una mezcla es un conjunto de k distribuciones probabilísticas que representan k grupos, donde cada distribución ofrece la probabilidad de que una instancia en particular tenga un cierto conjunto de valores de atributos, si se conoce que es un miembro de ese grupo. El problema principal de EM es el desconocimiento tanto de la distribución de cada instancia de entrenamiento, como de los parámetros del modelo de mezclas. Por esa razón, EM adopta el mismo procedimiento utilizado para el algoritmo de agrupamiento K-Means, de esta forma inicia con algunas estimaciones iniciales para los parámetros del modelo de mezclas, y las utiliza para calcular las probabilidades de los grupos para cada instancia. Finalmente, utiliza estas probabilidades para volver a estimar los parámetros.

c) Farthest-First Traversal

El algoritmo farthest first comienza seleccionando aleatoriamente una instancia que pasa a ser el centro (centroide) del grupo. Se calcula la distancia entre cada una de las instancias y el centro. La instancia que se encuentre más alejada del centro más cercano es seleccionada como el nuevo centro del grupo. Este proceso se repite hasta alcanzar el número de grupos buscado.

2.2. Modelo de espacio vectorial

En el modelo del espacio vectorial los documentos de texto son representados mediante vectores de términos. Si las palabras son elegidas para ser términos entonces la dimensionalidad del vector es tan grande como el vocabulario del corpus. Si un término ocurre en el documento, su valor en el vector es distinto de 0. La contribución de cada término en la representación

de cada documento es calculada mediante un esquema de pesado de términos, el cual se basa en simples estadísticas de términos. Una opción común para el pesado de los términos, es el factor de frecuencia de términos (tf) y el factor de frecuencia inversa de los documentos (idf).

A continuación definimos las formulas para *tf-idf* (en inglés, term frequency - inverse document frecuency).

tf es la cantidad de apariciones del termino k_i en el documento d_j:

$$tf_{i,dj} = c(k_i, d_j) \tag{2.2}$$

El valor de tf nos indica qué tan importante es un término t dentro de un documento d_j.

idf expresa que la importancia de un término es inversamente proporcional con el número de documentos en los que el término aparece. La medida se define como:

$$idf(t_i) = log\frac{N}{n_i} \tag{2.3}$$

Donde se supone que hay N documentos en la colección y que el término t_i ocurre n_i veces en ellos [6].

2.3. Indización aleatoria

La indización aleatoria (RI), es una metodología del espacio vectorial, cuya idea básica es acumular vectores de contexto de cada una de las palabras basándose en la co-ocurrencia de los datos. Lo primero que realiza esta técnica es asignar un vector índice, el cual es único y aleatorio a cada contexto a nivel documento o a nivel palabra. El vector indice está formado por cantidades iguales de +1 o -1. Esta técnica es inherentemente incremental y no requiere una fase de reducción de dimensión [11].

La técnica de indización aleatoria se puede describir en los dos siguientes pasos:

1. A cada contexto en los datos 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. Los vectores de contexto se producen mediante el recorrido del texto, y cada vez que una palabra ocurre en un contexto (en un documento o dentro de una ventana deslizante de contexto), esta d-dimensión de contexto del vector índice es agregada al vector de contexto para la palabra en cuestión. Las palabras son representadas por vectores de contexto d-dimensionales que son efectivamente la suma de los contextos de las palabras.

2.4. Representaciones holográficas reducidas

Las Representaciones Holográficas Reducidas (en inglés, Holographic Reduced Representations, HRR) son vectores n-dimensionales que representan estructura textual mediante representaciones distribuidas. La operación que se emplea para construir las asociaciones entre vectores es la convolución circular. Para representar a una asociación se crea un vector de la misma dimensionalidad que la de los vectores asociados, de esta forma, en la convolución circular no hay problema con respecto al crecimiento de los vectores y puede ser usada para sistemas conexionistas con vectores de ancho fijo, además de que preserva la similitud estructural.

La convolución circular $\otimes$ mapea dos vectores n-dimensionales en un vector $\mathbf{z}$. Si $\mathbf{x}$ y $\mathbf{y}$ son vectores n-dimensionales (subíndice 0 hasta n-1), entonces los elementos de $\mathbf{z}=\mathbf{x}\otimes\mathbf{y}$ son definidos como:

$$Z_i = \sum_{k=0}^{n-1} x_k y_{i-k} \text{ } \textit{para i=0 a n-1 (los índices son módulo-n)} \tag{2.4}$$

La convolución circular puede ser vista como un operador de multiplicación de vectores y tiene muchas propiedades algebráicas en común con la

multiplicación escalar y de matriz. Es conmutativa, asociativa y bilineal, tiene un vector identidad y un vector cero, además de que para muchos vectores existe su correspondiente inversa.

Gracias a las propiedades de preservación de similitud de la convolución circular, las representaciones serán aún más similares si las entidades están envueltas en roles similares. Así resulta que la HRR puede reflejar tanto la similitud superficial como la estructural de una manera que claramente se recuperen analogías textuales de manera similar a como lo hacemos los humanos [13]. Si bien la HRR se ha utilizado para recuperar analogías [13], en clasificación y recuperación de información [3], hasta donde se tiene conocimiento, no se ha utilizado para realizar agrupamiento de documentos.

Capítulo 3

Trabajo Relacionado

En esta sección se mencionan algunos trabajos relacionados con agrupamiento de documentos.

Guan, et al. En [12] presentan un nuevo algoritmo de agrupamiento de texto llamado propagación de la afinidad mediante semillas (en inglés, Seeds Affinity Propagation, SAP), mediante la extensión del algoritmo propagación de la afinidad (AP) con una nueva medida de similitud asimétrica llamada TRI-SET, la cual captura información estructural de los textos, y con un nuevo enfoque de aprendizaje semi-supervisado, donde explotan el conocimiento de un pequeño conjunto de objetos etiquetados contra un gran número de objetos no etiquetados. También proponen una nueva forma de definir los valores iniciales del algoritmo, la cual es llamada "semillas".Dado que la medida de similitud juega un papel importante en el algoritmo AP, en el algoritmo SAP se introdujo una nueva medida de similitud, la cual consta de tres conjuntos de características llamados co-características (CFS) que es la intersección entre el conjunto de características del documento d_i y el documento d_j, similar a la ecuación del Coseno; características unilaterales (UFS) que toma en cuenta aquellas características que no comparte el documento d_i con en documento d_j; y co-características más significativas(SFC) las cuales podrían ser frases clave o las etiquetas asociadas a cada documento. Bajo este enfoque cada término en el texto se sigue considerando como una característica y cada documento como un vector. Sin embargo, no todas las características y los vectores son calculados al mismo tiempo, si no uno a la vez.

El algoritmo propuesto, emplea a las "semillas" como ejemplares para obtener el número exacto de grupos. El método de construcción de "semillas"

puede encontrar rapidamente las características representativas en los objetos etiquetados. Para analizar el comportamiento del nuevo algoritmo se utilizó el corpus Reuters-21578, y se realizó una comparación detallada con los siguientes cuatro algoritmos: k-Means, AP con coeficiente coseno (AP(CC)),(AP (Tri-set)), AP combinado con el nuevo método semi-supervisado de construcción de "semillas" y con el coeficiente coseno (SAP(CC)). Las métricas utilizadas para la evaluación de los resultados obtenidos, fueron F-measure y entropía.

Cleuziou en [5] presenta un nuevo enfoque para explorar el espacio de búsqueda de posibles cubiertas para recuperar una adecuada organización en los grupos traslapados (o cubiertas). Propone una nueva función objetivo para minimizar bajo restricciones de multi-asignación, es decir, que explora el espacio de cubiertas en vez del espacio de particiones como k-Means y un algoritmo asociado a este criterio, llamado OKM, el cual es una generalización del algoritmo K-Means. Menciona que la tarea de asignar cada dato a uno o varios grupos no es una tarea trivial, por lo cual propone una heurística que consiste en desplazarse a través de la lista de grupos prototipo desde el más cercano al más lejano, y asignando el vector x_i mientras su imagen $\phi(x_i)$ es mejorada, la nueva asignación es conservada si es mejor que la anterior, asegurándose de que la función del error cuadrado disminuya. 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 para proveer mejores grupos traslapados.

Zimmerling 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 utilizó los conjuntos de datos Classic3 y Reuters-21578, como métrica de evaluación se empleó F-measure. Los algoritmos propuestos no mejoraron substancialmente al algoritmo K-Means en cuanto a tiempo de ejecución y número de iteraciones, debido a la naturaleza de los datos.

Su, et al. En [16], muestran un nuevo algoritmo de agrupamiento híbrido, el cual se basa en el algoritmo de cuantificación 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 su problema con entrenamiento insuficiente, ya que se producen pocas ite-

raciones de entrenamiento por cada estado en su estructura dinámica. Los pesos de los nodos de salida del agrupamiento GSC son considerados como los vectores prototipo iniciales de VQ, después de varias ejecuciones de entrenamiento, los nuevos pesos de los vectores prototipo ganadores de VQ reemplazarán a los pesos de los nodos de GCS. Tal proceso puede considerarse como una fase de entrenamiento adicional en la agrupación de GCS que resuelve el problema de entrenamiento insuficiente. Se denota como D el conjunto de documentos y F como el 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 en resultados previos. Todos los documentos del corpus Reuters se normalizan usando *tf-idf*, se emplearon los algoritmos K-Means, VG,GCS y la propuesta híbrida. De acuerdo a los resultados de los experimentos, el método propuesto alcanza mejor desempeño que los otros métodos.

Kamel y Ayad en [1], proponen combinar agrupaciones producidas por diferentes técnicas de agrupamiento para descubrir los tópicos de los documentos de texto, la agregación de estas agrupaciones revela una mejor estructura 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*, se emplearon para la agrupación por agregación algoritmos jerárquicos, incrementales y particionales. Se utiliza F-measure para evaluar y comparar los tópicos extraídos y la calidad de la agrupación antes y después de la agregación. La evaluación experimental muestra 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.

A continuación se muestra una tabla con las características principales de cada trabajo relacionado:

Tabla 3.1: Trabajos relacionados.

Referencia	**Algoritmo**	**Métrica de similitud**	**Características**
Guan, et al. en [12]	SAP	Tri-set	Enfoque semi-supervisado. Método de construcción de "semillas".
Cleuziou en [5]	OKM	Error cuadrado	Grupos traslapados. Extensión de K-Means.
Zimmerling en [10]	KKz	Error cuadrado	Extensión de K-Means. Semillas aleatorias.
Su, et al. en [16]	Hibrido GCS-VQ	Gradiente descendente	Enfoque de redes neuronales artificiales. Empleo de VQ para mejorar las salidas de GCS.
Kamel y Ayad en [1]	Single-link, complete-link, Leader y K-Means	De acuerdo a los enfoques empleados	Agregación de agrupaciones producidas por técnicas jerárquicas, incrementales y particionales. Modelo de espacio vectorial.

Capítulo 4

Metodología

La presente investigación tiene como hipotesis que mediante el uso de HRR, la tarea de agrupamiento puede mejorar al incluir como atributos, en la descripción de los documentos, términos con más información que las palabras simples. El diagrama a bloques correspondiente a nuestra metodología se puede observar en la Figura 4.1.

Figura 4.1: Cada documento se preprocesó con el fin de obtener los documentos representados como HRRs y finalmente, aplicamos un algoritmo de agrupamiento para crear los clusters en que se separarán los documentos.

A continuación describimos cada paso de la metodología propuesta:

1. **Selección de documentos.** Se seleccionaron diferentes subconjuntos de la colección total de documentos, con el fin de comparar nuestros resultados con algunos de los trabajos relacionados.

2. **Preprocesamiento de documentos.** Se eliminaron símbolos de puntuación y las palabras vacías de cada uno de los documentos, se realizó el truncamiento de las palabras con PorterStemmer [9](algoritmo para eliminar las terminaciones morfológicas más comunes de las palabras en inglés) y se etiquetó cada uno de los términos con MontyLingua [7] implementada en Python, para obtener las categorías sintácticas de los términos. Los documentos preprocesados quedaron de la siguiente forma: <término>/<categoría>. De esta forma un documento con el siguiente texto: "COMPUTER TERMINAL SYSTEMS <CPML> COMPLETES SALE", despues del preprocesamiento quedó como:
comput/NNP
termin/NNP
system/NNP
cpml/NNP
complet/NNP
sale/NN
A continuación se muestra una lista de las categorías y sus significados.

Tabla 4.1: Lista de categorías sintácticas.

Cat.	**Significado**	**Cat.**	**Significado**	**Cat.**	**Significado**
CC	Coordinating conjunction	PDT	Predeterminer	CD	Cardinal number
POS	Possessive ending	DT	Determiner	PRP	Personal pronoun
EX	Existential there	PRP$	Possessive pronoun	FW	Foreign word
RB	Adverb	IN	Preposition or subordinating conjunction	JJ	Adjective
RBR	Adverb, comparative	JJR	Adjective, comparative	RBS	Adverb, superlative
JJS	Adjective, superlative	RP	Particle	LS	List item marker
SYM	Symbol	MD	Modal	TO	to
NN	Noun, singular or mass	UH	Interjection	NNS	Noun, plural
VB	Verb, base form	NNP	Proper noun, singular	VBD	Verb, past tense
NNPS	Proper noun, plural	VBG	Verb, gerund or present participle	VBN	Verb, past participle
WP	Wh-pronoun	VBP	Verb, non-3rd person singular present	WP$	Possessive wh-pronoun
VBZ	Verb, 3rd person singular present	WRB	Wh-adverb	WDT	Wh-determiner

3. **Generación de HRRs**. Para representar los documentos empleando las HRRs se siguieron los siguientes pasos:

a) Se generaron vectores de contexto para el vocabulario de la colección utilizando la indización aleatoria.

b) Se creó la representación HRR para cada término utilizando la convolución circular para relacionar el término con su categoría sintáctica.

c) Los vectores representativos (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.

4. **Agrupamiento de documentos.** Se emplearon los algoritmos K-Means, EM y FarthestFirst para el agrupamiento de los documentos de texto.

Capítulo 5

Experimentos

En este capítulo se describen los experimentos realizados en esta investigación, con el objetivo de evaluar las HRRs en agrupamiento. Comenzamos describiendo el corpus empleado para los experimentos y las métricas de evaluación empleadas. Posteriormente se especifican los subconjuntos del corpus seleccionado para las pruebas realizadas. También se detallan los pasos de cada experimento realizado. Por último, se especifica el software utilizado para los experimentos.

5.1. Corpus

Existe una gran variedad de corpus para el agrupamiento de documentos. Para evaluar el efecto de las HRRs en la tarea de agrupamiento, en esta tesis se utilizó el corpus Reuters-21578 [4], pues está constituido por noticias de diversos contextos que presentan diferentes características.

El corpus Reuters es una colección de 21,578 artículos financieros que aparecieron en el servicio de noticias Reuters en 1987, distribuido en 91 categorías. Cada documento fue manualmente clasificado por los editores del servicio de noticias. El número de documentos asignado a cada categoría varía, asignando a algunas categorías un gran número de documentos, como por ejemplo a la categoría *earn*, mientras que a otras categorías, como *rye*, muy pocos documentos.

5.2. Métricas de evaluación

Para evaluar el desempeño del agrupamiento, se aplicaron dos tipos de métricas, F-measure y Pureza, las cuales fueron empleadas para comparar los grupos generados con el conjunto de categorías creadas manualmente del corpus Reuters. F-measure es una combinación armónica de los valores de precisión y recuerdo. Se realiza el cálculo de la precisión (*P(i,j)*) y del recuerdo (*R(i,j)*) para cada grupo j y para cada clase i, basándose en los resultados de los algoritmos de agrupamiento y las clases pre-establecidas del corpus utilizado. Precisión y Recuerdo se definen a continuación:

Precisión: Es la fracción de un grupo que consiste de objetos de una clase en específico. La precisión de un grupo j con respecto a la clase i es:

$$P(i,j) = \frac{m_{ij}}{m_j} \tag{5.1}$$

Donde m_j es el número de objetos en el grupo y m_{ij} es el número de objetos de la clase i en el grupo j.

Recuerdo: Es el grado en el que un grupo contiene todos los objetos de una de una clase en específico. El recuerdo de un grupo j con respecto a la clase i es:

$$R(i,j) = \frac{m_{ij}}{m_i} \tag{5.2}$$

Donde m_{ij} se define de la misma forma que en el caso de la precisión y m_i es el número de objetos en la clase i.

F-measure es la combinación de ambas, precisión y recuerdo, que mide el grado en el cual un grupo contiene solamente objetos de una clase en particular y todos los objetos de esta clase. La F-measure de un grupo j con respecto a la clase i es:

$$F(i,j) = \frac{2 * P(i,j) * R(i,j)}{P(i,j) + R(i,j)} \tag{5.3}$$

La F-measure global para todo el resultado de agrupamiento es definida a continuación:

$$F = \sum_i \frac{N_i}{N} max_j F(i,j) \tag{5.4}$$

Donde N_i es el número de objetos de la clase i, N es el número total de documentos en el conjunto de datos y max_j es la máxima F-measure obtenida del grupo j.

Para calcular la Pureza cada grupo es asignado a la clase que es más frecuente en el grupo, de esta forma la precisión de esta asignación es medida contando el número de asignaciones correctas de documentos y dividiéndola entre N. Formalmente se define a continuación:

$$Pureza(\Omega, C) = \frac{1}{N} \sum_k max_j \left| w_k \bigcap c_j \right| \tag{5.5}$$

Donde Ω es el conjunto de los grupos y C es el conjunto de las clases. Se interpreta a w_k como el conjunto de documentos en Ω y a c_j como el conjunto de documentos en C.Un mal agrupamiento, tiene un valor de pureza cercano a cero, y un buen agrupamiento tiene una pureza cercana a uno.

5.3. Entorno experimental

Para el primer experimento, se ocuparon 68 clases, tanto para el conjunto de entrenamiento como el de prueba de Reuters, con un total de 9592 documentos. Posteriormente, con el fin de compararnos con los resultados reportados en [5], se utilizó un subconjunto de documentos distribuidos en 10

clases (*coffee, sugar, trade, rubber, earn, cpi, cotton, alum, bop* y *jobs*) para contar con un total de 3696 documentos únicos, tratando de aproximarnos lo más posible al conjunto de documentos utilizado en [5], hasta donde los detalles de dicha fuente nos permitió hacerlo. Ver tabla 5.1.

Tabla 5.1: Subconjunto de documentos considerados en los primeros experimentos.

Conjunto de datos	Documentos	Clases
total de documentos	9592	68
subconjunto	3696	10

Dado que el corpus Reuters tiene clases muy desbalanceadas, se empleó el experimento de cinco casos utilizado en [12]. 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 Figura 5.1.

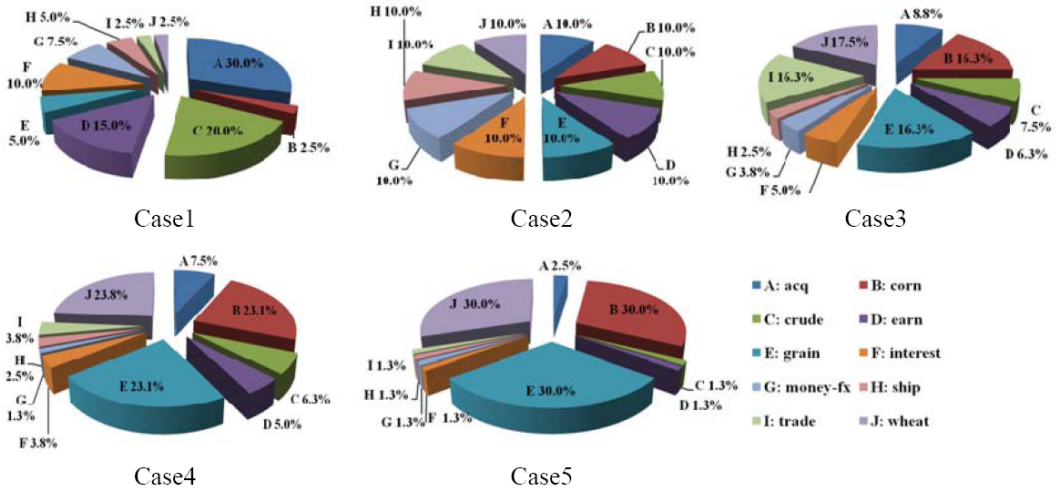

Figura 5.1: Porcentajes de documentos de cada una de las 10 clases seleccionadas como en [12]

En Reuters los documentos de las clases *corn, grain* y *wheat* son muy similares entre sí, debido a que comparten varios sinonimos entre sus términos

de *maiz*, *grano* y *trigo*, lo cual hace que sean muy difíciles de distinguir. Entonces la distribución de estas clases puede influir profundamente en los resultados de agrupamiento. Para cada uno de los cinco casos las clases *corn, grain* y *wheat* contienen los siguientes porcentajes de dificultad 10 % , 30 %, 50 %, 70 % y 90 %, respectivamente. Ver tabla 5.2, las clases mencionadas se presentan en otro color.

Tabla 5.2: Porcentajes de documentos por cada clase entre los 5 casos descritos en [12]

	% **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

Una vez realizado el preprocesamiento de los documentos, como en los pasos mencionados en metodología, se utilizó la indización aleatoria para generar la representación de las HRRs, tanto con una dimensión de 1024 como con una de 2048 componentes, con el fin de definir cuál es la dimensión adecuada para los experimentos posteriores.

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 los documentos y se eligieron los siguientes algoritmos: K-Means, EM y FarthestFirst.

Capítulo 6

Resultados

En este capítulo se presentan los resultados obtenidos con los conjuntos de datos seleccionados de acuerdo a como se especificó en el capítulo 4. Para cada experimento se muestra la medida de evaluación F-measure comparada con los resultados de la bibliografía.

El objetivo del primer experimento, fue definir la dimensionalidad de los vectores representativos para los experimentos posteriores. En la tabla 6.1 se muestran los resultados de las HRRs con dimensiones de 1024 y 2048 componentes. Podría esperarse que una representación de mayor dimensión tuviera un mejor desempeño al agrupar documentos, sin embargo, se puede observar que la representación con una dimensión de 1024 tiene una diferencia favorable con respecto a la de 2048 componentes, en términos de F-measure.

Tabla 6.1: Resultados de agrupamiento con dimensiones de 1024 y 2048 componentes

HRR			
Total de documentos	Clases	F-measure 1024	F-measure 2048
9592	68	0.3149	0.3086

En el segundo experimento se usaron las implementaciones de tres algoritmos de agrupamiento para definir el algoritmo con mejor desempeño. A continuación se muestran los resultados obtenidos para los algoritmos de agrupamiento k-Means, EM y FarthestFirst. En la Tabla 6.2, se puede observar que el algoritmo con peor desempeño, en cuanto a F-measure fue EM, y que aparentemente el algoritmo con mejor resultado fue FarthestFirst, sin

embargo, debido a que fue el peor en cuanto a Pureza, nos da un indicio para realizar un segundo análisis reduciendo el número de documentos. En la Tabla 6.4 se puede observar que los experimentos realizados con FarthestFirst confirman que su desempeño se vio afectado al emplear pocos documentos, lo que lo muestra como un algoritmo no robusto ante cambios en el número de documentos.

Tabla 6.2: Resultados de tres algoritmos de agrupamiento

Total de documentos	Clases	Algoritmo	F-measure 1024	Pureza
9592	68	K-Means	0.3149	0.7488
		EM	0.2621	0.7028
		FarthestFirst	0.4924	0.6015

En la Tabla 6.3 se presentan los resultados obtenidos con las HRRs de los conjuntos de documentos descritos en el capítulo 4 comparados con los resultados alcanzados por los diferentes métodos reportados en la bibliografía, hasta donde fue posible aproximar el número de clases y documentos. Puede verse que aparentemente las HRRs no mejoran los resultados del agrupamiento, presumiblemente porque nuestras pruebas se realizaron con cerca del triple de documentos que en el trabajo de Cleuziou en [5] empleando las siguientes 10 clases (*coffee, sugar, trade, rubber, earn, cpi, cotton, alum, bop* y *jobs*), y también con la diferencia de clases de la colección con 9592 documentos con respecto a los otros métodos comparados.

El objetivo de los siguientes experimentos fue evaluar la robustez de la representación propuesta con el algoritmo K-Means, empleando pequeños conjuntos de documentos. Los experimentos se realizaron con los cinco subconjuntos de 800 documentos que homogenizaron al conjunto de pruebas en el trabajo de Guan R. et al [12]. En la Tabla 6.4 se presenta una comparación de los resultados obtenidos, al aplicar las HRRs para representar documentos y el algoritmo K-Means para agruparlos, con los obtenidos en dicho trabajo. Puede observarse que la combinación de las HRRs con K-Means se desempeñó mejor que el algoritmo AP cuando se empleó el coeficiente coseno (CC) como métrica de similitud, e incluso que el algoritmo AP con la métrica de similitud Tri-set propuesta en [12]. También mejoró al algoritmo SAP cuando se utiliza el coeficiente Coseno como métrica de similitud. Por otro lado, dicho algoritmo combinado con la métrica Tri-set mejora a los resultados obtenidos con las HRR, mejora que con mucha probabilidad se debe a la diferencia en

Tabla 6.3: Resultados obtenidos de las diferentes propuestas de agrupamiento

Trabajo	Resultado reportado			HRR		
	Docs.	**Clases**	**F-measure**	**Docs.**	**Clases**	**F-measure**
An extended version of the k-Means method for overlapping clustering [5]	1308	10	0.76	3696	10	0.364
Topic discovery from text using aggregation of different clustering methods [1]	3000	10	.398	3696	10	0.364
An empirical study of k-Means initialization methods for document clustering [10]	8193	65	0.35	9592	68	0.314
Document clustering based on vector quantization and growing-cellstructure [16]	10,794	100	0.34	9592	68	0.314

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 modelo de espacio vectorial en un 48.63 % en promedio para los cinco casos considerados; en un 20.89 % 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 %, considerando que emplean métricas de similitud distintas.

Tabla 6.4: Comparación de resultados con el trabajo relacionado, utilizando F-measure

Referencia	**Algoritmo**	**CASO 1**	**CASO 2**	**CASO 3**	**CASO 4**	**CASO 5**	**PROM.**
HRR-1024	**K-MEANS**	0.5711	0.5671	0.6248	0.5421	0.4148	**0.5440**
	FartherstFirst	0.6049	0.3804	0.4643	0.5173	0.5298	0.4993
Text clustering with seeds affinity propagation	**SAP**	0.749	0.606	0.573	0.544	0.489	**0.592**
	SAP (CC)	0.662	0.519	0.511	0.450	0.385	0.505
	AP(Tri-Set)	0.577	0.482	0.419	0.364	0.290	0.426
	AP(CC)	0.450	0.450	0.450	0.450	0.450	0.450
	K-MEANS	0.518	0.397	0.368	0.280	0.269	**0.366**

A continuación se muestran los tiempos de agrupamiento con la representación propuesta.

Tabla 6.5: Tiempos de arupamiento con HRR y K-Means

Referencia	**Algoritmo**	**CASO 1**	**CASO 2**	**CASO 3**	**CASO 4**	**CASO 5**	**PROM.**
HRR	K-Means	16 seg	21 seg	8 seg	9 seg	7 seg	12.2 seg

Capítulo 7

Conclusiones

7.1. Principales resultados

En esta tesis se emplearon para representar documentos a las HRRs capturando información sintáctica de los mismos con la ayuda de la Indización Aleatoria. Los resultados de esta tesis reflejan que las HRRs de dimensión 1024 componentes tienen mayor efectividad en la tarea de agrupamiento con los subconjuntos seleccionados, en contraste con la representación con VSM, la cual tiene una dimensionalidad proporcional al tamaño del vocabulario del corpus. Los resultados se muestran competitivos en relación con los de los métodos reportados en los trabajos relacionados. Para los experimentos con 800 documentos se calculó una matriz de dimensión reducida (800×1024 componentes) teniendo un tiempo promedio de 0.20 minutos para el agrupamiento. Un efecto positivo de la reducción de dimensión es que se logra reducir el tiempo de ejecución del algoritmo de agrupamiento.

7.2. Trabajo futuro

El trabajo que se sugiere que puede realizarse más adelante es el siguiente:

- Emplear diferentes métricas de evaluación de agrupamiento para definir con precisión el beneficio de las HRRs en el agrupamiento de documentos.

- Realizar pruebas adicionales con diferentes colecciones de documentos.

- Emplear un corpus en lenguaje español para corroborar la eficiencia de la representación propuesta.
- Realizar un análisis cualitativo de la utilidad de la información sintáctica en la discriminación entre grupos, dado que aporta mayor información que el indizado tradicional en VSM.

Apéndice A

Publicación derivada de la tesis

Norma L. Cuautle-Rivera, Maya Carrillo, A. López-López, Agrupación de Documentos Utilizando Representaciones Holográficas Reducidas, 4to Congreso Mexicano de Inteligencia Artificial (COMIA 2012),(por publicarse).

Bibliografía

[1] Kamel M. Ayad H. Topic discovery from text using aggregation of different clustering method. In *Proceedings of the15th Conference of the Canadian Society for Computational Studies of Intelligence on Advances in Artificial Intelligence Springer-Verlag London, UK*, 2012.

[2] Dubes R. C. How many clusters are best?-an experiment. In *Pattern Recogn. 20s*, pages 6645–663, 1987.

[3] Eliasmith C. Carrillo M. and Lopez-Lopez A. Combining text vector representations for information retrieval. In *In: V. Matousek, P. Mautner (eds) Text, Speech and Dialogue. Proceedings of the 12th International Conference Text, Speech and Dialogue, LNAI*, volume 5729, pages 24–31, 2009.

[4] Lewis D.D. Reuters-21578 text categorization test collection. 2004.

[5] Cleuziou G.. An extended version of the k-means method for overlapping clustering. In *LIFO - University of Orleans*, 2008.

[6] Salton G. and Buckley C. Term-weighting approaches in automatic text retrieval. In *Information Processing and Management 24 513 523*, 1988.

[7] Liu H. Montylingua: An end-to-end natural language processor with common sense. In *MIT Media Lab*, 2004.

[8] Jain A. K. and Dubes R. C. Algorithms for clustering data. In *Prentice Hall College Div*, page 320, 1988.

[9] Porter M. The porter stemming algorithm. 1980.

[10] Zimmerling M. An empirical study of k-means initialization methods for document clustering. In *Term Paper, Dresden University of Technology (Germany)*, 2008.

[11] Magnus Sahlgren. An introduction to random indexing. In *In: Methods and Appli-cations of Semantic Indexing Workshop at the 7th International Conference on Terminology and Knowledge Engineering, Copenhagen, Denmark*, 2005.

[12] Guan R. Yang C. Shi X., Marchese M. and Liang Y. Text clustering with seeds affinity propagation. In *IEEE Trans. Knowledge and Data Eng.*, volume 23, pages 627–637, 2011.

[13] Plate T.A. Analogy retrieval and processing with distributed vector representation. In *Technical report CS-TR-98-4 Victoria University of Wellington, Computer Science. 16 p. Longer version of an invited submission to the Workshop on Advances in Analogy Research held at New Bulgarian University, Sofia, Bulgaria*, 1998.

[14] Plate T.A. Distributed representation in: Encyclopedia of cognitive science. In *Macmillan Reference Ltd*, 2002.

[15] Plate T.A. Holographic reduced representation, distributed representation for cognitive structures. In *CSLI Publications*, 2003.

[16] Su Z. Zhang L., Pan Y. Document clustering based on vector quantization and growing-cell structure. In *Developments in 16th international conference on industrial and engineering applications of artificial intelligence and expert systems, Loughborough,UK*, pages 326–336, 2003.