Mineración de Textos Científicos

Sheila Maricela Pinto Cáceres

Universidade Estadual de Campinas

IC - UNICAMP

Índice

- Introducción
- Colección
- Preprocesamiento
  - Pasos a seguir
- Clusterización
- Resultados
- Conclusiones

Introducción

- Gran crecimiento de información, mayor cantidad de documentos en formato digital.
- Procesar estos datos y agruparlos automáticamente se torna en una tarea realmente necesaria.
- En este trabajo se describe el proceso para agrupar textos en una colección considerablemente grande dando énfasis a la parte de preprocesamiento de la colección.

Colección

- Artículos ACM → 2003 – 2007.
- Idioma →Inglés
- Numero de documentos → 75 171 artículos

Ano	Quantidade
2003	13145
2004	17028
2005	19177
2006	18337
2007	7484
Total	75171

Pasos

1. Documento -> palabras

Estandarización de la colección

- Se analiza la data, verificando que los datos sean representativos y que no estén dañados o faltantes. Se observo que de los 75 171 paper:
  - 2875 de estos artículos no tienen abstract.
  - Total → 72806
- Formatos no usables requieren ser convertidos a texto plano. Se extrajo cada abstract de la colección desde el portal de la ACM (html → txt).
- Minusculas.

Pasos

Pasos

1. Documento -> palabras
2. Palabras -> raíz (stemming)

   Ej: ‘walk’, ‘walking’, ‘walks’ →’walk’

Stemmización

- Extracción de las raíces de palabras.
- El proceso de stemmización esta basado en el algoritmo propuesto por Porter que anteriormente ha sido adaptado para Ingles, Español y portugues.

Recuperación de Texto (Review)

Pasos

1. Documento -> palabras
2. Palabras -> raíz (stemming)

   Ej: ‘walk’, ‘walking’, ‘walks’ →’walk’
3. Stop List: Palabras comunes son eliminadas

   Ej: ‘El’, ‘un’, etc.

Ejemplo:

"…Representation, detection and learning are the main issues that need to be tackled in designing a visual system for recognizing object. categories …"

Ejemplo:

represent ***detect*** ***learn***

~~Representation, detection~~ and ~~learning~~ ~~are the~~

main ***issue*** ***tackle*** ***design***

~~main issues~~ that ***need*** to be ~~tackled in designing~~

visual system ***recognize*** ***category***

~~a visual system for recognizing~~ ***object*** ~~categories~~.

…

Pasos

Pasos

1. Documento -> palabras
2. Palabras -> raíz (stemming)
   Ej: ‘walk’, ‘walking’, ‘walks’ →’walk’
3. Stop List: Palabras comunes son eliminadas
   Ej: ‘El’, ‘un’, etc.
4. Creación del “bag of words”.
   Cada palabra -> identificador único

Word	**ID**
represent	1
detect	2
learn	3
……	….

Pasos

Pasos

1. Documento -> palabras
2. Palabras -> raíz (stemming)

   Ej: ‘walk’, ‘walking’, ‘walks’ →’walk’
3. Stop List: Palabras comunes son eliminadas

   Ej: ‘El’, ‘un’, etc.
4. Creación del “bag of words”. Cada palabra -> identificador único
5. Cada documento-> Vector de k componentes

Frecuencia

- Se hicieron pruebas con tres tipos de frecuencias en una palabra a ser utilizadas sobre la matriz de datos:
- 1. Booleano: Si el atributo aparece o no en el texto.
- 2. Frecuencia: cantidad de veces que un atributo aparece en un texto.
- 3. Frecuencia relativizada: utilizando term frequency - inverse document frequency.

Documento ->k-vector de f(palabra)

- Vocabulario de K palabras.
- Documento -> vector de k componentes
- Documento $V_d=(t_1, \ldots, t_i, \ldots, t_k)$
  (0,0, … 3,… 4,…. 5, 0,0)

Estándar Weighting
Term frecuency – Inverse document frecuency tf-idf

$$t_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i}$$

Word Frecuency

Inverse Document Frecuency

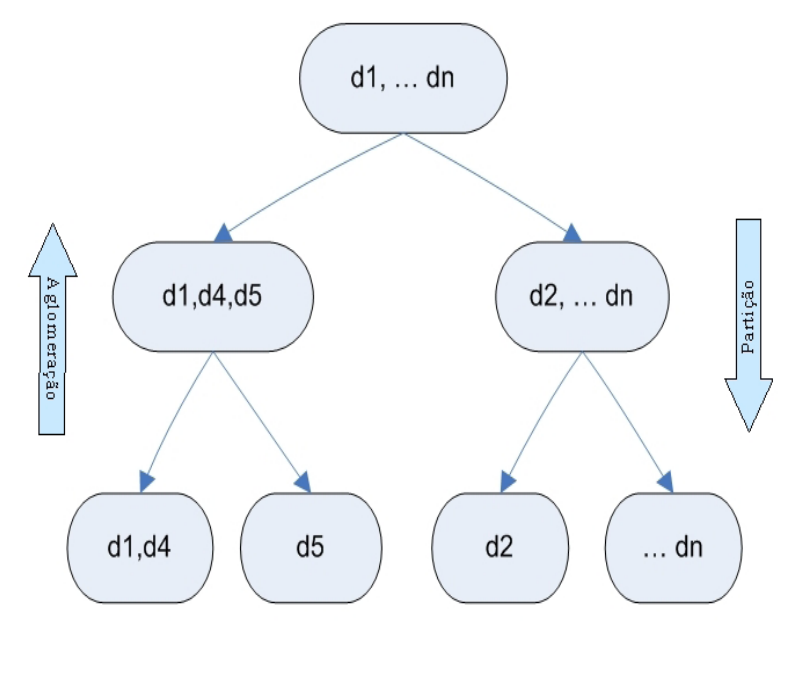
n_{id}: numero de ocurrencias de la palabra i en documento d
n_d: numero de palabras en el documento d
n_i: numero de ocurrencias del termino i en la bd
N: numero de documentos en la bd

Selección de atributos

- A pesar de la stemmización y el empleo de stop list, quedaron →101 000 palabras (dimensiones)
- Como primera prueba o primer escenario se uso el conjunto total de dimensiones, tomando 8 horas de processamiento para generar una matriz de 14 GB de texto.
- Para la reducción de dimensionalidades se usaron los cortes de Luhn[6]
  - Se establecen dos puntos de corte para los atributos de acuerdo con la frecuencia de los términos en un determinado documento mediante un histograma de frecuencias ordenado e forma descendente. Adoptando como puntos de corte los dos puntos de inflexión de la curva.(puntos no exactos)

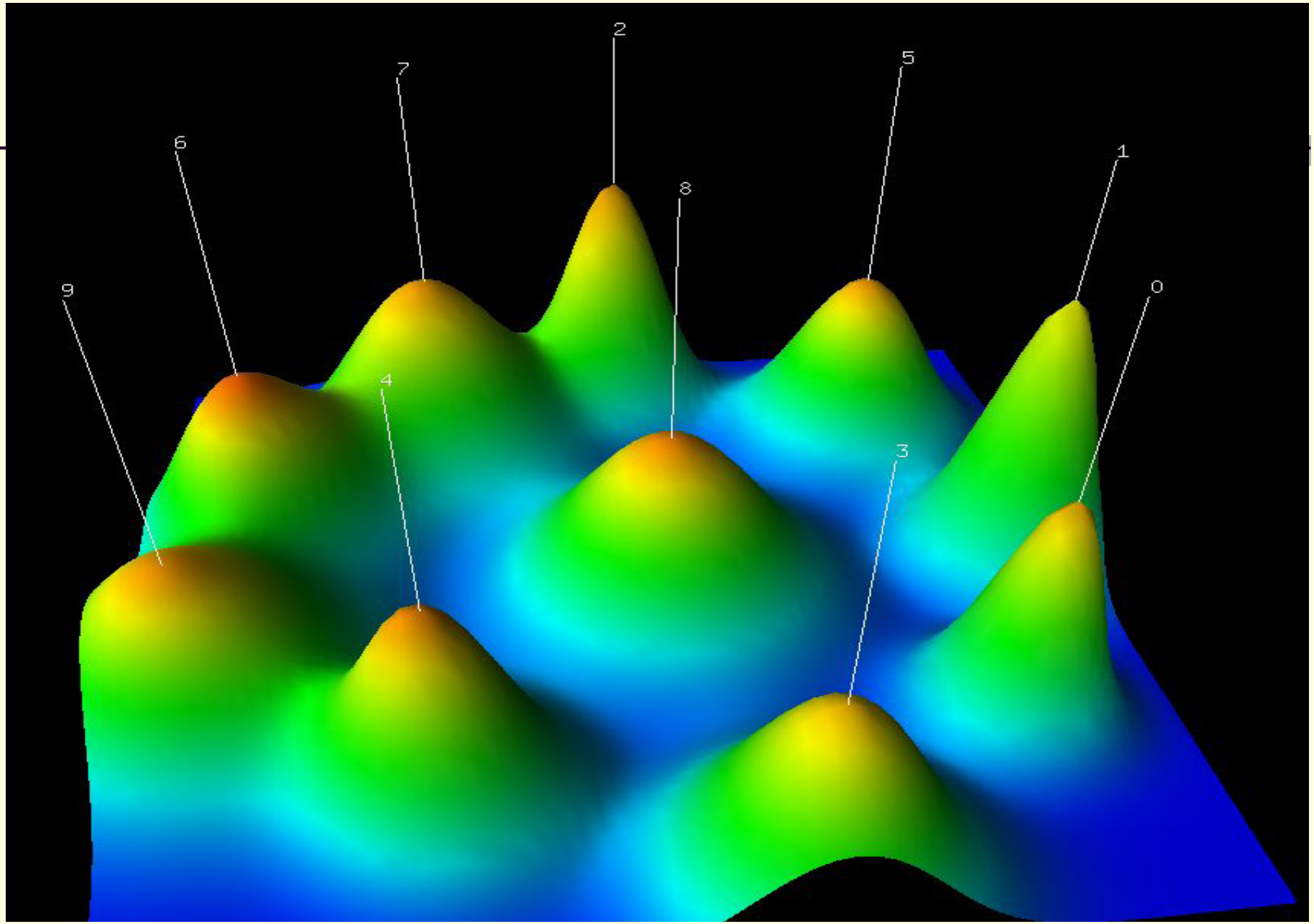
Mineración

- Clusterización jerarquica: construye una jerarquia o un árbol de clusters conocido como dendograma mostrando como son agrupados los documentos. Cada cluster en el árbol contiene diferentes clusteres hijos permitiendo diferente numero de clusters



Experimentos y Resultados

- Fueron montados 8 escenarios diferentes.
- Todos los escenarios fueron analisados de acordo com a similaridade interna e a similaridade externa dos clusters.
- Para aplicar algoritmos de clusterización en 3D se uso la herramienta gCluto.
- Cada pico es uma curva gaussiana y es uma estimativa de la distribucion de los datos de cada cluster.
  - A altura de cada pico → similaridad interna de cada cluster.
  - volume → cantidad de documentos.
  - Color del pico → Proporcional al desviación interna del cluster.
- Se describen tambien los siguientes datos.
  - Size: N´umero de textos no cluster
  - ISim: Similaridade interna
  - ESim: Similaridade externa


2
7
5
6
1
8
9
0
4
3

Experimentos y Resultados

- Cluster 0, Size: 3212, ISim: 0.033, ESim: 0.016
  approxim, polynomial, linear, matrix, equat, curv
- Cluster 1, Size: 4589, ISim: 0.040, ESim: 0.016
  equat numer, equat numer discret, equat boundari, equat numer differenti, equat nonlinear
- Cluster 2, Size: 4582, ISim: 0.036, ESim: 0.018
  memori parallel, processor hardwar, benchmark, processor instruct, memori cach
- Cluster 3, Size: 7496, ISim: 0.029, ESim: 0.014
  prove, bound give, polynomial, vertic edg, theorem
- Cluster 4, Size: 5242, ISim: 0.031, ESim: 0.017
  product, decis, busi, organ, market price
- Cluster 5, Size: 6537, ISim: 0.027, ESim: 0.016
  circuit, low, voltag current, circuit gate, circuit delai
- Cluster 6, Size: 7489, ISim: 0.026, ESim: 0.017
  semant, defin, express, logic, extens
- Cluster 7, Size: 10699, ISim: 0.029, ESim: 0.017
  protocol, traffic, packet, internet, wireless
- Cluster 8, Size: 10699, ISim: 0.027, ESim: 0.017
  extract, retriev, databas, classif classifi, train
- Cluster 9, Size: 12234, ISim: 0.025, ESim: 0.015
  student, interfac, project, collabor, creat

Experimentos y Resultados

```
--------------------------------------------------------------------------------
Hierarchical Tree...
--------------------------------------------------------------------------------
                      Size   XSim      Gain
18                   [72779, 2.09e-11, +1.84e-02]
|---------13          [15297, 6.34e-10, +3.71e-02]
|         |---------3  [ 7496, 0.00e+00, +0.00e+00]
|         |-----10     [ 7801, 3.09e-09, +4.56e-02]
|               |---0  [ 3212, 0.00e+00, +0.00e+00]
|               |---1  [ 4589, 0.00e+00, +0.00e+00]
|-17                   [57482, 3.21e-11, +2.50e-02]
  |-16                 [21818, 2.59e-10, +2.59e-02]
  | |---------------5   [ 6537, 0.00e+00, +0.00e+00]
  | |-15               [15281, 5.72e-10, +2.80e-02]
  |   |-------------2   [ 4582, 0.00e+00, +0.00e+00]
  |   |-------------7   [10699, 0.00e+00, +0.00e+00]
  |-----14             [35664, 1.32e-10, +3.53e-02]
        |---12         [24965, 2.98e-10, +3.90e-02]
        |   |-------6  [ 7489, 0.00e+00, +0.00e+00]
        |   |-11       [17476, 6.54e-10, +4.19e-02]
        |     |-----4  [ 5242, 0.00e+00, +0.00e+00]
        |     |-----9  [12234, 0.00e+00, +0.00e+00]
        |-----------8  [10699, 0.00e+00, +0.00e+00]
```

Escenario 6 con 10 clusters

#	Size	ISim	ESim
0	3212	0.033	0.016
1	4589	0.040	0.016
2	4582	0.036	0.018
3	7496	0.029	0.014
4	5242	0.031	0.017
5	6537	0.027	0.016
6	7489	0.026	0.017
7	10699	0.029	0.017
8	10699	0.027	0.017
9	12234	0.025	0.015

Conclusiones

- La selección de atributos tuvo un papel trascendental antes de aplicar los métodos de clusterización sobre los datos directamente, ya que de tener todos los datos las funciones o hiperplanos de separación de los planos nos podrían llevar a un overfitting.
- Los cortes de Luhn sirvieron eficientemente en la reduccion de dimensionalidad.
- La clusterizacoin de todos los artículos demostro que una colección de esta magnitud envuelve una mayor dificultad en la definicion de los clusters, no entanto con un conjunto menor de articulos estos fueron mejor definidos despues de aplicar el algoritmo de clusterizacion
- La eleccion del numero de clusters deseados demostró tambien tener un papel importante en la clasificación de cada cluster.

Trabajos Futuros

- Adicionar Key-words, autor o titulo en la información del documento.
- Usar bigramas y/o trigramas

Referencias

- [1] A. Asuncion and D. Newman. UCI machine learning repository, 2007.
- [2] S. Banerjee and T. Pedersen. The design, implementation, and use of the ngram statis- tics package. In A. F. Gelbukh, editor, CICLing, volume 2588 of Lecture Notes in Computer Science, pages 370–381. Springer, 2003.
- [3] I. Dhillon, J. Kogan, and C. Nicholas. Feature selection and document clustering. In M. W. Berry, editor, Survey of Text Mining, pages 73–100. Springer, 2003.
- [4] L. Liu, J. Kang, J. Yu, and Z. Wang. A comparative study on unsupervised feature selection methods for text clustering. Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE ’05. Proceedings of 2005 IEEE International Conference on, pages 597–601, 30 Oct.-1 Nov. 2005.
- [5] Ding, C., He, X., Zha, H., Gu, M., and Simon, H. (2001). Spectral min-max cut for graph partitioning and data clustering. In Proc. 1st IEEE International Conference on Data Mining, pages 107–114.
- [6] Luhn, H. P. (1958). The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2).
- [7] Porter, M. F. (1980). An algorithm for suffix stripping. Program, 14(3):130–137.

Gracias