INSTITUTO DE INGENIERÍA UNAM

Métodos para la obtención automática de términos en un área de especialidad

Ing. Alberto Barrón Cedeño
Instituto de Ingeniería
Posgrado en Ciencia e Ingeniería de la Computación
UNAM
alberto@pumas.ii.unam.mx

Programa

- Introducción
- Técnicas para la extracción de términos
- El algoritmo C-value/NC-value
  - Parte lingüística de C-value
  - Parte estadística de C-value
  - NC-value
- Un breve ejemplo
- Conclusiones

Introducción

Proyecto CONACyT:
Extracción de conceptos en textos de especialidad a través del reconocimiento de patrones lingüísticos y metalingüísticos

Creación de diccionarios
Creación de tesauros y ontologías

Desarrollo de buscadores
Clasificadores de documentos

Técnicas para la extracción de términos

Técnicas lingüísticas

- Los términos siguen ciertos patrones sintácticos.
- Se requiere etiquetado POS.

Ejemplos: Lexter, Heid

Técnicas para la extracción de términos

Técnicas estadísticas

- Un suceso que ocurre frecuentemente es relevante

Ejemplo: ANA

Técnicas para la extracción de términos

- Técnicas híbridas
  - Combinan las dos técnicas anteriores

Ejemplos: Acabit, TermoStat, Termine

C-value/NC-value

Es un método híbrido para la extracción automática de términos multipalabra en inglés.

En particular, se ha desarrollado para el procesamiento de documentos del área de biomedicina.

C-value/NC-value

C-value (única etapa híbrida)

lingüística

- Etiquetado de partes de la oración
- Detección de candidatos a término
- Eliminación de candidatos por medio de una lista de paro

estadística

- Ordenación de candidatos con base en su frecuencia y longitud

Texto original

Tal y como sale de fábrica el disco duro no puede ser utilizado por un sistema operativo.

Antes tenemos que definir en él una o más particiones y luego hemos de darles un formato que pueda ser entendido por nuestro sistema.

Texto etiquetado

Tal_tal/QU y_y/CC como_como/CSUBX sale_salar/VLfin de_de/PDE fábrica_fábrica/NC el_el/ART disco_disco/NC duro_duro/ADJ no_no/NEG puede_poder/VMfin ser_ser/VSinf utilizado_utilizar/VLadj por_por/PREP un_un/ART sistema_sistema/NC operativo_operativo/ADJ ._./FS

Antes_antes/ADV tenemos_tener/VLfin que_que/CQUE definir_definir/VLinf en_en/PREP él_él/PPX una_un/ART o_o/CC más_más/ADV particiones_particiones/NC y_y/CC luego_luego/CSUBF hemos_haber/VLfin de_de/PDE darles_dar/VLinf un_un/ART formato_formato/NC que_que/CQUE pueda_poder/VMfin ser_ser/VSinf entendido_entender/VLadj por_por/PREP nuestro_nuestro/PPO sistema_sistema/NC ._./FS

C-value, etapa lingüística: detección de candidatos

<NC | NP | PE>$^{+}$

servidor, tarjeta madre, MB, arquitectura von Neumann

<NC> <ADJ> (<PDE> <NC | NP>)*

sistema operativo, unidad central de procesamiento

<NC> <PDE> <NC | NP | NMEA>

ángulo de visión, ancho de banda, tasa de transferencia

C-value, etapa lingüística: detección de candidatos

<VLFIN | VLINF>

compilar, descifrar, ensamblar

<NC>? <ACRNM>

DDR2, IP, slot ISA

<NC> <PDE> ((<NC> <ADJ>) | (<ADJ> <NC>))

computadora de uso general

C-value, etapa lingüística: detección de candidatos

Tal y como sale de fábrica el disco duro no puede ser utilizado por un sistema operativo.

Antes tenemos que definir una o más particiones en el disco duro y formatearlo para que pueda ser entendido por nuestro sistema.

C-value, etapa lingüística: lista de paro

Más de 200 palabras (nombres y adjetivos) que no se espera que aparezcan dentro de los términos del área.

Ej. (computación): detalle, elegante, importante, mayoría, mercado, opción, tamaño, vez

C-value, etapa lingüística: aplicando la lista de paro

~~**fábrica**~~

disco duro

sistema operativo

partición

formato

sistema

Aspectos considerados

1. Frecuencia total de ocurrencia del sintagma en el corpus
2. Frecuencia total de ocurrencia del sintagma como parte de sintagmas más largos
3. Número de dichos candidatos de mayor longitud
4. Longitud del candidato a término

C-value, etapa estadística: cálculo de C-value

$$C\text{-}value = \begin{cases} (1 + log_2|a|) * f(a) & si\ a\ no\ aparece\ en\ otros\ candidatos \\ (1 + log_2|a|) * \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right) & de\ lo\ contrario \end{cases}$$

a es el sintagma candidato

$|a|$ es la longitud de a

$f(a)$ es la frecuencia de ocurrencia de a en el corpus

T_a es el conjunto de candidatos de mayor longitud que contienen a a

$P(T_a)$ *es el* número de esos candidatos (incluye al mismo candidato)

$\Sigma f(b)$ *es la* ocurrencia total de a como subcadena del sintagma candidato b tal que $|a|<|b|$

C-value, etapa estadística: cálculo de C-value

candidato	longitud	frecuencia
disco duro	2	3
sistema operativo	2	2
sistema	1	2
partición	1	1

C-value, etapa estadística: cálculo de C-value

candidato	**longitud**	**frecuencia**
sistema operativo	2	2

$$C\text{-}value = (1 + log_2|a|) * f(a)$$

$$C\text{-}value(sistema\ operativo) = (1 + log_2(2)) * 2$$

$$C\text{-}value(sistema\ operativo) = 4$$

C-value, etapa estadística: cálculo de C-value

candidato	**longitud**	**frecuencia**
sistema	1	2

$$C\text{-}value = (1 + log_2|a|) * \left(f(a) - \frac{1}{P(T_a)} \sum_{b \in T_a} f(b) \right)$$

$$C\text{-}value(sistema) = (1 + log_2(1)) * \left(2 - \frac{1}{2} * 2 \right)$$

$$C\text{-}value(sistema) = 2$$

C-value, etapa estadística: cálculo de C-value

candidato	longitud	frecuencia
disco duro	2	3
sistema operativo	2	2
sistema	1	2
partición	1	1

C-value, etapa estadística: cálculo de C-value

candidato	longitud	frecuencia	C-value
disco duro	2	3	6
sistema operativo	2	2	4
partición	1	1	1
sistema	1	2	1

El contexto en el que se hallan los candidatos es importante para ellos.

Las palabras que actúan con los términos no suelen ser arbitrarias.

Son **palabras de contexto** aquellos nombres, adjetivos y verbos que aparecen en el contexto de un candidato.

NC-value: el contexto del candidato

Un llavero USB es un **pequeño dispositivo** de almacenamiento que utiliza la memoria flash para guardar la información sin necesidad de pilas.

Tienen una capacidad de almacenamiento que va desde algunos megabytes hasta 8 gigabytes.

NC-value: el contexto del candidato

Un **llavero USB** es un **pequeño dispositivo de almacenamiento** que **utiliza** la **memoria flash** para guardar la información sin necesidad de pilas.

Tienen una capacidad de almacenamiento que va desde algunos megabytes hasta 8 gigabytes.

NC-value: el contexto del candidato

Un llavero USB es un pequeño dispositivo de **almacenamiento** que **utiliza** la memoria flash para **guardar** la **información** sin necesidad de pilas.

Tienen una capacidad de almacenamiento que va desde algunos megabytes hasta 8 gigabytes.

NC-value: el contexto del candidato

Un llavero USB es un pequeño dispositivo de almacenamiento que utiliza la memoria flash para guardar la información sin necesidad de pilas.

Tienen una capacidad de almacenamiento que **va** desde algunos **megabytes** hasta 8 gigabytes.

NC-value: el contexto del candidato

Un llavero USB es un pequeño dispositivo de almacenamiento que utiliza la memoria flash para guardar la información sin necesidad de pilas.

Tienen una capacidad de **almacenamiento** que va desde algunos megabytes hasta 8 **gigabytes**.

NC-value: el contexto del candidato

Un llavero USB es un pequeño dispositivo de almacenamiento que utiliza la memoria flash para guardar la información sin necesidad de pilas.

Tienen una capacidad de almacenamiento que va desde algunos **megabytes** hasta 8 gigabytes.

NC-value, palabras de contexto

candidato	**palabras de contexto**
disco duro	formatear, guardar, GB
sistema operativo	instalar, configurar, cargar
partición	disco duro, crear, swap
sistema	levantar, iniciar, apagar

NC-value, palabras de contexto

$$weight(w) = \frac{t(w)}{n}$$

w es la palabra de contexto analizada

weight(w) es el peso asignado a la palabra *w*

t(w) es el número de candidatos con los que aparece la palabra *w*

n es el número total de candidatos considerados (para expresarlo como una probabilidad)

NC-value, palabras de contexto

"es necesario **formatear** el *disco* y ***crear*** dos **particiones**"

"el tipo de **partición** para el sistema operativo **Linux es ext3** "

$$weight(formatear) = \frac{1}{2} \qquad weight(particion) = 1$$

NC-value: determinando el contexto

Tal y como **sale** de **fábrica** el **disco duro** no **puede ser utilizado** por un sistema operativo.

Antes tenemos que definir una o más particiones en el disco duro y formatearlo para que pueda ser entendido por nuestro sistema.

NC-value: determinando el contexto

Tal y como **sale** de **fábrica** el **disco duro** no **puede ser utilizado** por un sistema operativo.

Antes tenemos que definir una o más particiones en el disco duro y formatearlo para que pueda ser entendido por nuestro sistema.

NC-value: cálculo de NC-value

$$NC\text{-}value(a) = 0.8C\text{-}value(a) + 0.2\sum_{b \in C_a} f_a(b)weight(b)$$

$$NC\text{-}value(disco\ duro) = 0.8(6) + 0.2\left(1 * 1 + 1 * \frac{1}{2}\right)$$

$$NC\text{-}value(disco\ duro) = 5$$

NC-value: cálculo de NC-value

candidato	frecuencia	C-value	NC-value
disco duro	3	6	5.3
sistema operativo	2	4	5.0
formato	2	2	1.8
partición	1	1	1.4
sistema	2	1	0.8

C-value/NC-value: comparación (posición)

candidato	frecuencia	C-value	NC-value
usuario	1	1	1
estación de trabajo	69	7	8
problema	119	140	138
memoria flash	176	85	82

C-value/NC-value muestra de resultados

USB ElBus de Serie Universal(USB, de sussiglaseninglés Universal Serial Bus) es uninterfazque provee unestándar de bus seriepara conectardispositivosa unordenador personal(generalmente a unPC). Unsistema USBtiene undiseño asimétrico, que consiste en un soloservidory múltiplesdispositivosconectados en unaestructura de árbolutilizandoconcentradores especiales. Se pueden conectar hasta 127dispositivosa un soloservidor, pero lasumadebe incluir a losconcentradorestambién, así que eltotal de dispositivosrealmente usables es algo menor. Fue desarrollado afinalesde 1996 por sieteempresas:IBM,Intel,Northern Telecom,Compaq,Microsoft,Digital Equipment CorporationyNEC. Elestándarincluye latransmisión de energía

Conclusiones

El algoritmo C-value/NC-value ha mostrado ser una buena opción en búsqueda de la extracción de términos en el español.

Los errores pueden ser “soportados” si se considera que una persona puede tardar unas 3 horas en obtener la terminología de un documento de 2381 palabras.

Se tarda unas horas en obtener la terminología de un corpus de 140,000.

Conclusiones

INSTITUTO DE INGENIERÍA UNAM

¡Gracias!

Alberto Barrón Cedeño

alberto@pumas.ii.unam.mx