# Capítulo 2. Marco Teórico

## 2.1 Métodos para la Generación del Resumen Automático

Resumir un documento significa reducir a términos breves y precisos lo esencial de un asunto o materia, Mani indica, que la confección del resumen de una fuente, consiste en extraer y presentar al usuario el contenido más importante, condensado y adaptado a las necesidades de la aplicación o del usuario [Mani, 2001].

Existen diversos estudios que muestran la complejidad al crear un resumen de manera automática [Sparck-Jones, 2006]; los principales inconvenientes que se tienen son los siguientes:

- No se conoce la estructura del documento.
- El documento no contiene el tema o la categoría a la que pertenece.
- El tipo de documento influye al momento de hacer el resumen, el resumen de un libro lo podemos obtener por la tabla de contenido del mismo, en cambio el resumen de una noticia lo podemos obtener por medio del título.
- Al realizar un resumen es importante dominar el tema para poder lograr la comprensión del mismo.
- Para realizar un resumen se puede tomar en cuenta una o varias fuentes, en caso de trabajar con varios documentos se debe tomar en cuenta la redundancia de la información, así como también, los cambios que ha sufrido la información a través del tiempo.

Dentro del procesamiento del lenguaje natural se ha trabajado desde hace mucho tiempo con diferentes técnicas para generar un resumen de forma automática; actualmente las

principales técnicas se clasifican en dos categorías: basadas en conocimiento y basadas en selección [Jurafsky & Martin, 2000].

### 2.1.1 Basadas en Conocimiento

Las técnicas basadas en conocimiento realizan un análisis semántico al documento que se va a resumir, después se realiza una transformación para obtener la representación semántica de la información esencial del documento y por último se genera el resumen en lenguaje natural [Jurafsky & Martin, 2000]. Como podemos observar se necesita mucho conocimiento para poder utilizar esta técnica y por lo tanto va a tener un gran peso el dominio del documento.

Dentro de esta categoría se encuentran las técnicas que hacen uso de plantillas. El objetivo de esta técnica es extraer los datos relevantes del documento, los cuales servirán para llenar una plantilla o un formulario y finalmente se generará el resumen usando el lenguaje natural [Sparck-Jones, 2006]. Los inconvenientes que observamos en la técnica de plantillas son dos; el primero es que los documentos deben ser del mismo género, debido a que se debe realizar el diseño de la plantilla; el segundo inconveniente es que no se toman en cuenta los datos que no están incluidos en la plantilla a pesar de ser relevantes.

### 2.1.2 Basadas en Selección

El objetivo de esta técnica es seleccionar las oraciones más importantes del documento y presentarlas como su resumen. La ponderación de las oraciones se basa en diferentes métricas [Mateo *et al*, 2006] como son:

- Frecuencia de aparición de palabras
- Palabras indicativas

- Título del documento
- Nombres propios
- Tipografía del texto
- Posición de la oración dentro del documento

La ponderación basada en la frecuencia de aparición de palabras, toma las ideas de Luhn con respecto a que las palabras que aparecen frecuentemente dentro de un documento son relevantes, mientras que las palabras con alta o baja frecuencia no son relevantes.

La ponderación basada en palabras indicativas busca las oraciones que contienen palabras como “importante”, “esencial”, “para concluir”, etc., y les otorga un mayor peso; de esta forma estas oraciones estarán presentes en el resumen del documento. Los inconvenientes de esta métrica son dos; el primero es que se debe crear una lista de las palabras indicativas y estas palabras dependen del género del documento; el segundo es que la lista de palabras indicativas no es independiente del idioma del documento.

La ponderación basada en el título del documento, otorga mayor peso a las oraciones que contienen las palabras del título debido a que, en algunos casos, indica el contenido del documento.

La ponderación basada en nombres propios, otorga mayor peso a las oraciones que los contienen, de esta forma se proporciona mayor información en el resumen del documento.

La ponderación basada en tipografía del texto, otorga mayor peso a las oraciones que tienen texto resaltado por un formato específico, como puede ser, negritas, tamaño de letra, subrayado y mayúsculas.

La ponderación basada en la posición de la oración dentro del documento, otorga mayor peso a las oraciones de los primeros y los últimos párrafos, debido a que estos contienen

más información relevante que los párrafos centrales. El inconveniente que presenta esta métrica es que la suposición que hace, depende del género del documento.

Algunos trabajos además de usar algunas de estas métricas usan un corpus perteneciente a un dominio y también usan técnicas de aprendizaje.

La técnica basada en selección es muy sencilla de aplicar, ya que no necesita un gran número de recursos como las técnicas basadas en conocimiento; otra ventaja que presenta es que se puede aplicar a cualquier tipo de documento; sin embargo, los problemas que presenta esta técnica son dos, pérdida de coherencia y desequilibrio en el resumen [Mateo *et al*, 2006]. La pérdida de coherencia en el resumen se genera debido a la selección de las oraciones del documento, ya que probablemente se contarán con anáforas, las cuales hacen referencia a oraciones que no están presentes en el resumen. La solución a este problema consiste en resolver la anáfora, debemos aclarar que la solución de la anáfora no es un problema trivial. El desequilibrio del resumen se refiere a que no se incluyan algunas oraciones que contienen información relevante del documento.

Observando las características de las técnicas para generar un resumen automático, se decidió usar la técnica basada en selección, debido a que los documentos que se van a resumir son páginas Web, las cuales pertenecen a diferentes géneros con diversos dominios y esta técnica es independiente de estos conceptos; otra razón por la que se tomó esta decisión, es que el sistema se usará en tiempo real y la velocidad del sistema es un factor muy importante, y esta técnica usa una menor cantidad de recursos, comparada con las técnicas basadas en conocimiento.

Para poder encontrar las palabras clave de un documento podemos usar la técnica del punto de transición, y si además deseamos encontrar los términos multipalabra del documento podemos usar el concepto de información mutua.

## 2.2 Técnica del Punto de Transición

El concepto del punto de transición surgió por las observaciones de George K. Zipf, el cual observó que el ser humano, al realizar una tarea, trata de minimizar la pérdida de energía y en caso de tener diferentes opciones, las personas optarían por las que impliquen el menor esfuerzo; en base a esto, Zipf formuló la ley de frecuencias de un texto, en la que establece que las palabras con mayor frecuencia son las usadas comúnmente como pueden ser los artículos, preposiciones, etc., mientras que las palabras con baja frecuencia son poco relevantes dentro del documento, las palabras que tienen una frecuencia media son las que representan la información del documento [Urbizagástegui, 2006]. El punto de transición es la frecuencia de una palabra del documento, la cual hace una división entre las palabras de baja frecuencia y las palabras de alta frecuencia. Los términos más cercanos al punto de transición, los podemos considerar como las palabras claves que representan al documento.

La fórmula para encontrar el punto de transición es la siguiente:

$$PT \quad \frac{\sqrt{1 \quad 8I_1} \quad 1}{2}$$

en donde $I_1$ representa el número de palabras con frecuencia 1 [Jiménez *et al*, 2006].

En diferentes trabajos que se han desarrollado podemos observar que tomando una banda de frecuencias desde un 15% hasta un 25% alrededor del punto de transición podemos contar con las palabras clave que representan al documento.

Cuando se trabaja con documentos pequeños surge el problema de que no se tienen términos cuya frecuencia esté alrededor del punto de transición. En estos casos, el punto de transición se busca por el método de inspección, el cual toma los términos con la frecuencia más baja que no se repite, estos términos serán las palabras clave que representen al documento.

## 2.3 Términos Multipalabra

Las palabras clave sirven para representar un documento, pero también los términos multipalabra podrían ser palabras importantes dentro del documento, este concepto también es conocido por el nombre de colocaciones.

Vilares y Ribalas definen al concepto de colocación de la siguiente manera, dos palabras forman una colocación si aparecen regularmente en una lengua y el significado de la aparición conjunta es diferente a la simple suma de significados de las palabras individuales [Vilares & Ribadas, 2006].

Existen diferentes métodos para identificar los términos multipalabra, entre los principales métodos se encuentran: el método de bigramas, el cual está orientado a la búsqueda de combinaciones de palabras sustentándose en la medida de su información mutua, y también podemos mencionar los métodos que hacen un análisis sintáctico [Gelbukh & Sidorov, 2006], los cuales establecen patrones sintácticos (Adjetivo-Sustantivo, Sustantivo-Adjetivo, etc.) para poder reconocer los términos multipalabra.

### 2.3.1 Información Mutua

La información mutua representa la medida de asociación entre las palabras. Según Fano, si dos puntos (palabras) '*x*' y '*y*', tienen probabilidades P(*x*) y P(*y*), entonces su información mutua, IM(*x*,*y*), se define como [Fano, 1961]:

$$IM(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}$$

Informalmente, la información mutua compara la probabilidad de observar '*x*' y '*y*' de manera conjunta (la probabilidad de unión) con las probabilidades de observar '*x*' y '*y*' independientemente. Si hay una asociación genuina entre '*x*' y '*y*', entonces la probabilidad común P(*x*,*y*) será mucho más grande que P(*x*) P(*y*), y por lo tanto I(*x*,*y*) > 0. Si no existe una relación de interés entre '*x*' y '*y*', entonces P(*x*,*y*) ≈ P(*x*) P(*y*) y entonces I(*x*,*y*) ≈ 0. Si '*x*' y '*y*' están en distribución complementaria, entonces P(*x*,*y*) será mucho menor que P(*x*) P(*y*), forzando que I(*x*,*y*) < 0.

En nuestro caso, las probabilidades P(*x*) y P(*y*) son estimadas contando la frecuencia de ocurrencia de '*x*' y de '*y*' en un texto. Las probabilidades comunes, P(*x*,*y*), son estimadas contando el número de veces que '*x*' es seguido por '*y*' en una ventana de '*w*' palabras. Posteriormente, se aplica la función normalizando entre el tamaño del documento (*N*).

Se ha observado que el cociente de la asociación llega a ser inestable cuando los valores de conteo de frecuencia son muy pequeños, por lo que regularmente se usa un umbral de 5 para la frecuencia de ocurrencia conjunta. Esta aproximación es totalmente arbitraria, sin embargo, en la práctica se toma como un valor útil [Salazar *et al*, 2006].

En nuestro trabajo, hemos tomado en cuenta el hecho de que los documentos analizados son heterogéneos en cuanto al tamaño, por lo que se ha decidido normalizar la función para el cálculo de información mutua. También se han detectado ciertas posibilidades de que la

frecuencia de ocurrencia conjunta de términos sea cero. En este último caso, la fórmula planteada con anterioridad arrojaría un valor de infinito. Finalmente, la fórmula ha quedado expresada como sigue:

$$IM(x,y) \quad \log_2 \frac{N \quad fr(x,y)}{fr(x)\,fr(y)} \quad 1$$

en donde *fr*(*x*) y *fr*(*y*) es la frecuencia de ocurrencia del término '*x*' y '*y*', respectivamente, y *fr*(*x*,*y*) es la frecuencia de ocurrencia conjunta de los términos '*x*' y '*y*' [Salazar *et al*, 2006].

Tomando en cuenta los conceptos mencionados anteriormente, se escogió el método basado en selección para encontrar el resumen de un documento. Las métricas que se van a tomar en cuenta son cuatro:

1. Palabras clave
2. Términos multipalabra
3. Título del documento
4. Tipografía del documento

Se van a proporcionar dos combinaciones diferentes de métricas; la primera va a tomar en cuenta las palabras clave, los términos multipalabra, el título y la tipografía del documento; la segunda combinación tomará en cuenta las palabras clave y los términos multipalabra. Se tomaron en cuenta estas dos opciones para poder comparar los resultados y de esta manera utilizar la mejor combinación.

## 2.4 Uso del paquete javax.swing.text.html

El paquete javax.swing.text.html tiene las clases e interfaces necesarias para trabajar con documentos HTML; por medio de este paquete, podemos analizar documentos HTML para conocer la información del documento. Esta clase permite definir métodos que son llamados cada vez que se está analizando una etiqueta; entre estos métodos se encuentran:

public void handleEndTag(HTML.Tag tag, int position)

public void handleText(char[] text, int position)

public void handleStartTag(HTML.Tag t, MutableAttributeSet a, int position)

El método handleStartTag se usa para analizar la etiqueta de apertura, el método handleEndTag sirve para manejar la etiqueta de cierre y por último tenemos al método handleText para manejar el texto que se encuentra entre las etiquetas. Esta información se obtuvo del sitio http://java.sun.com/j2se/1.4.2/docs/api/