Extracción de Información Semántica a Partir de Categorías de Texto

Estado del Arte

Andrés Romero Rodríguez

Maestría en Ingeniería de Sistemas y Computación
Universidad Nacional de Colombia
Bogotá D.C.

June 9, 2006

Contenido

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Introducción

- Grandes volúmenes de información
  - Documentos de trabajo
  - Información corporativa
  - Correo electrónico
- Dificultad para su manejo
- Problemas para acceder a la información relevante
- Obtención de información
  - Palabras clave
  - Descripción
  - Redes semánticas
  - Ontologías

Contenido

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Obtención de Información

- Se busca obtener información adicional de cada una de las categorías de los documentos
- Es posible contar con conocimiento previo acerca del contexto
- Esta información extra puede responder preguntas como:
  - Por qué se generan determinados grupos?
  - Como se relacionan las categorías entre si?
  - Por qué un documento pertenece a una categoría?
  - Qué tipo de información hay en una categoría?

Fases en el Manejo del Conocimiento

1. Obtención del conocimiento
2. Organización y estructuración del conocimiento
3. Refinamiento del conocimiento
4. Distribución del conocimiento

Fases en el Manejo del Conocimiento

1. Obtención del conocimiento
2. Organización y estructuración del conocimiento
3. Refinamiento del conocimiento
4. Distribución del conocimiento

Fases en el Manejo del Conocimiento

1. Obtención del conocimiento
2. Organización y estructuración del conocimiento
3. Refinamiento del conocimiento
4. Distribución del conocimiento

Fases en el Manejo del Conocimiento

1. Obtención del conocimiento
2. Organización y estructuración del conocimiento
3. Refinamiento del conocimiento
4. Distribución del conocimiento

Contenido

Red Semántica

- Representan conocimiento en patrones de nodos y arcos interconectados
- Dan estructura a los conceptos como una red
- Los nodos de la red representan conceptos
- Los arcos representan relaciones entre estos conceptos

Red Semántica

- Representan conocimiento en patrones de nodos y arcos interconectados
- Dan estructura a los conceptos como una red
- Los nodos de la red representan conceptos
- Los arcos representan relaciones entre estos conceptos

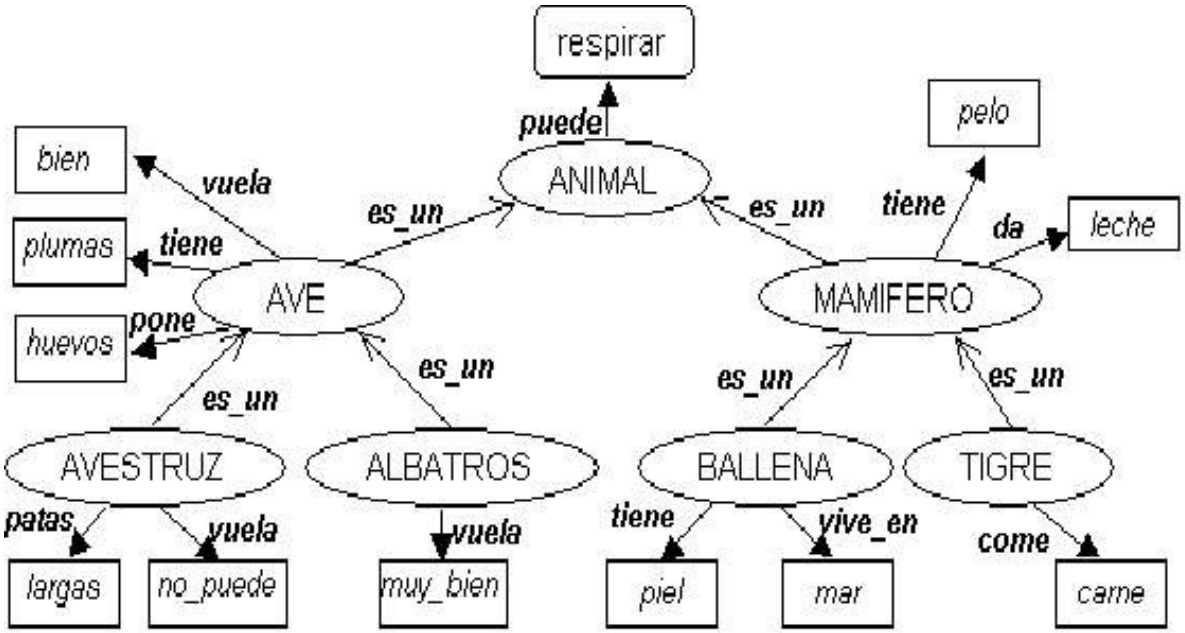
Red Semántica

- Representan conocimiento en patrones de nodos y arcos interconectados
- Dan estructura a los conceptos como una red
- Los nodos de la red representan conceptos
- Los arcos representan relaciones entre estos conceptos

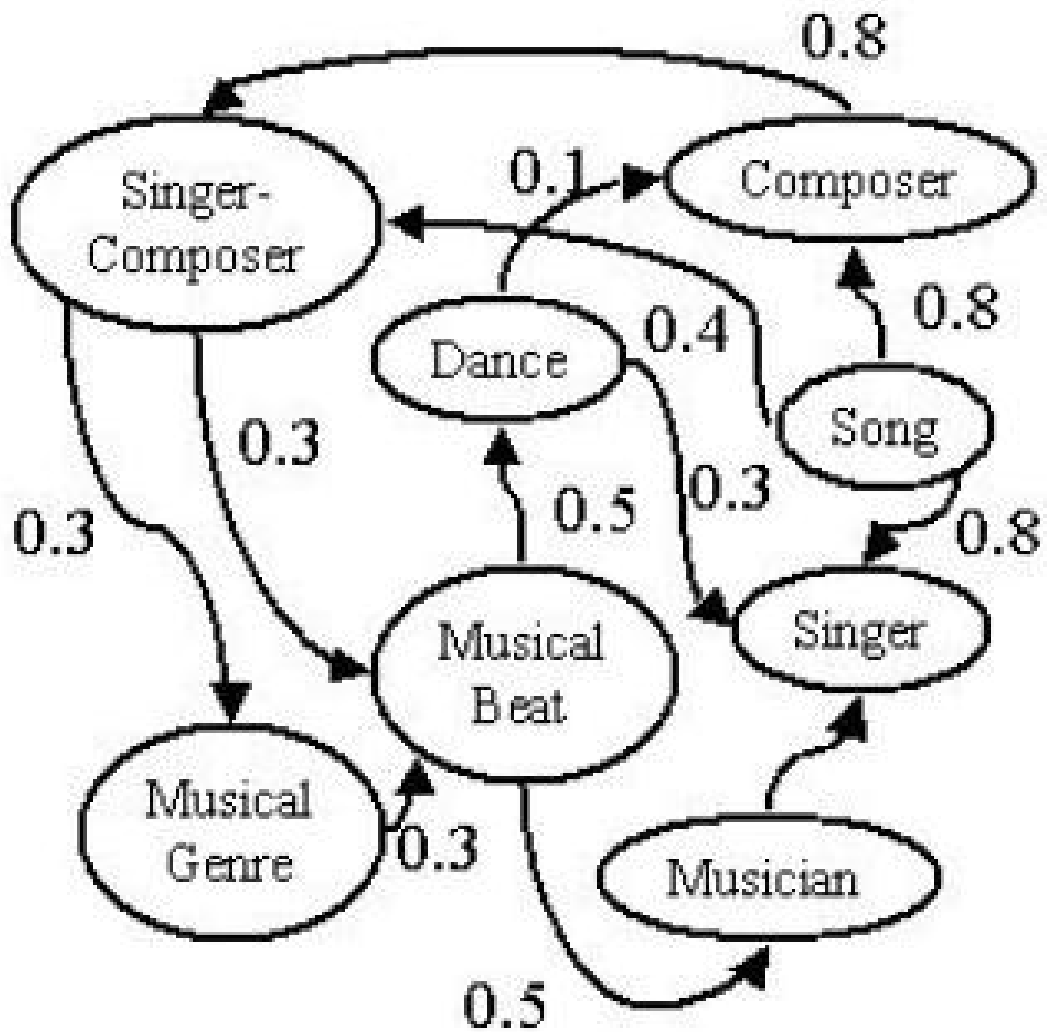
Red Semántica

- Representan conocimiento en patrones de nodos y arcos interconectados
- Dan estructura a los conceptos como una red
- Los nodos de la red representan conceptos
- Los arcos representan relaciones entre estos conceptos

Redes Semánticas



Redes Semánticas



Contenido

Ontologías

- Se entiende como un entendimiento común y compartido de un dominio
- Se construye preferiblemente como un esfuerzo colaborativo de expertos en el dominio.
- Adicionan contenido semántico a documentos web

Ontologías

- Se entiende como un entendimiento común y compartido de un dominio
- Se construye preferiblemente como un esfuerzo colaborativo de expertos en el dominio.
- Adicionan contenido semántico a documentos web

Ontologías

- Se entiende como un entendimiento común y compartido de un dominio
- Se construye preferiblemente como un esfuerzo colaborativo de expertos en el dominio.
- Adicionan contenido semántico a documentos web

Extracción de Características

- Se debe contar con una representación adecuada de los documentos.
- La mayoría de técnicas se basan en un conjunto de palabras que representan el contenido almacenado en un documento.
- La diferencia entre estos esquemas radica en el valor que se almacena para cada palabra.
- Existen otras técnicas de representación poco usadas.

Extracción de Características

- Se debe contar con una representación adecuada de los documentos.
- La mayoría de técnicas se basan en un conjunto de palabras que representan el contenido almacenado en un documento.
- La diferencia entre estos esquemas radica en el valor que se almacena para cada palabra.
- Existen otras técnicas de representación poco usadas.

Extracción de Características

- Se debe contar con una representación adecuada de los documentos.
- La mayoría de técnicas se basan en un conjunto de palabras que representan el contenido almacenado en un documento.
- La diferencia entre estos esquemas radica en el valor que se almacena para cada palabra.
- Existen otras técnicas de representación poco usadas.

Extracción de Características

- Se debe contar con una representación adecuada de los documentos.
- La mayoría de técnicas se basan en un conjunto de palabras que representan el contenido almacenado en un documento.
- La diferencia entre estos esquemas radica en el valor que se almacena para cada palabra.
- Existen otras técnicas de representación poco usadas.

Contenido

Vectores de Palabras

- La representación del conjunto de documentos se hace mediante una matriz A.

$$A = (a_{ik})$$

- a_{ik} representa el peso de la palabra i en el documento k.

Vectores de Palabras

- La representación del conjunto de documentos se hace mediante una matriz A.

$$A = (a_{ik})$$

- a_{ik} representa el peso de la palabra i en el documento k.

Boolean Weighting

$$a_{ik} = \begin{cases} 1 & \text{si } f_{ik} > 0 \\ 0 & \text{en otro caso} \end{cases}$$

Word Frequency Weighting

$$a_{ik} = f_{ik}$$

tf $\times$ idf Weighting

$$a_{ik} = f_{ik} * \log \frac{N}{n_i}$$

tfc Weighting

$$a_{ik} = \frac{f_{ik} * \log \frac{N}{n_i}}{\sqrt{\sum_{j=1}^{M} [f_{ik} * \log \frac{N}{n_i}]^2}}$$

ltc Weighting

$$a_{ik} = \frac{\log\left(f_{ik} + 1.0\right) * \log\frac{N}{n_i}}{\sqrt{\sum_{j=1}^{M}\left[\log\left(f_{ik} + 1.0\right) * \log\frac{N}{n_i}\right]^2}}$$

Entropy Weighting

$$a_{ik} = \log(f_{ik} + 1.0) * (1 + \frac{1}{\log(N)} \sum_{j=1}^{N} [\frac{f_{ij}}{n_i} \log \frac{f_{ij}}{n_i}])$$

Contenido

n-grams

- Consiste en utilizar secuencias de palabras seleccionadas.
- Se utilizan técnicas estadísticas como χ^2

n-grams

- Consiste en utilizar secuencias de palabras seleccionadas.
- Se utilizan técnicas estadísticas como χ^2

Part-Of-Speech Tagging

- Se le asigna a cada palabra del documento una categoría sintáctica.
- Se debe decidir cual es la mejor categoría dentro del contexto del documento.
- La misma palabra puede tener diferentes categorizaciones de acuerdo a su utilización.

Part-Of-Speech Tagging

- Se le asigna a cada palabra del documento una categoría sintáctica.
- Se debe decidir cual es la mejor categoría dentro del contexto del documento.
- La misma palabra puede tener diferentes categorizaciones de acuerdo a su utilización.

Part-Of-Speech Tagging

- Se le asigna a cada palabra del documento una categoría sintáctica.
- Se debe decidir cual es la mejor categoría dentro del contexto del documento.
- La misma palabra puede tener diferentes categorizaciones de acuerdo a su utilización.

Noun Phrases

- Frases que en conjunto tienen un mismo significado.
- Las palabras individuales no representan información relevante.
- Técnicas conocidas como *Named Entity Recognition*

Noun Phrases

- Frases que en conjunto tienen un mismo significado.
- Las palabras individuales no representan información relevante.
- Técnicas conocidas como *Named Entity Recognition*

Noun Phrases

- Frases que en conjunto tienen un mismo significado.
- Las palabras individuales no representan información relevante.
- Técnicas conocidas como *Named Entity Recognition*

Rocchio

- Método utilizado tradicionalmente.
- Construye un vector prototipo para cada categoría.
- Los documentos se clasifican calculando la distancia al vector prototipo.
- El vector prototipo es el promedio de todos los vectores que pertenecen a dicha categoría.

Rocchio

- Método utilizado tradicionalmente.
- Construye un vector prototipo para cada categoría.
- Los documentos se clasifican calculando la distancia al vector prototipo.
- El vector prototipo es el promedio de todos los vectores que pertenecen a dicha categoría.

Rocchio

- Método utilizado tradicionalmente.
- Construye un vector prototipo para cada categoría.
- Los documentos se clasifican calculando la distancia al vector prototipo.
- El vector prototipo es el promedio de todos los vectores que pertenecen a dicha categoría.

Rocchio

- Método utilizado tradicionalmente.
- Construye un vector prototipo para cada categoría.
- Los documentos se clasifican calculando la distancia al vector prototipo.
- El vector prototipo es el promedio de todos los vectores que pertenecen a dicha categoría.

Clasificadores Bayesianos

- Utiliza los datos de entrenamiento para calcular la probabilidad de cada categoría dados los valores de las características.
- Para calcular la probabilidad de tener una categoría dado un documento de prueba, se utiliza el teorema de Bayes:

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}$$

Clasificadores Bayesianos

- Utiliza los datos de entrenamiento para calcular la probabilidad de cada categoría dados los valores de las características.
- Para calcular la probabilidad de tener una categoría dado un documento de prueba, se utiliza el teorema de Bayes:

$$P(c_j|d) = \frac{P(c_j)P(d|c_j)}{P(d)}$$

K Nearest Neighbour

- Cada vector de entrada es comparado con los demás documentos.
- Se encuentran los K vecinos mas cercanos.
- Se utiliza un sistema de votación para determinar la categoría del documento de entrada.

K Nearest Neighbour

- Cada vector de entrada es comparado con los demás documentos.
- Se encuentran los K vecinos mas cercanos.
- Se utiliza un sistema de votación para determinar la categoría del documento de entrada.

K Nearest Neighbour

- Cada vector de entrada es comparado con los demás documentos.
- Se encuentran los K vecinos mas cercanos.
- Se utiliza un sistema de votación para determinar la categoría del documento de entrada.

Support Vector Machines

- Se mide el margen de separación de los datos de diferentes categorías.
- Realiza una reducción de la dimensionalidad junto con la tarea de clasificación.
- Los problemas con múltiples categorías deben ser tratados como varios problemas de clasificación binaria.

Support Vector Machines

- Se mide el margen de separación de los datos de diferentes categorías.
- Realiza una reducción de la dimensionalidad junto con la tarea de clasificación.
- Los problemas con múltiples categorías deben ser tratados como varios problemas de clasificación binaria.

Support Vector Machines

- Se mide el margen de separación de los datos de diferentes categorías.
- Realiza una reducción de la dimensionalidad junto con la tarea de clasificación.
- Los problemas con múltiples categorías deben ser tratados como varios problemas de clasificación binaria.

Preguntas Interesantes

- De que forma se deben representar los documentos de modo que no se pierda la estructura y las relaciones entre palabras y conceptos?
- Qué tipo de información se debe extraer de las categorías que sea representativa?
- Cómo representar la información extraida?

Preguntas Interesantes

- De que forma se deben representar los documentos de modo que no se pierda la estructura y las relaciones entre palabras y conceptos?
- Qué tipo de información se debe extraer de las categorías que sea representativa?
- Cómo representar la información extraida?

Preguntas Interesantes

- De que forma se deben representar los documentos de modo que no se pierda la estructura y las relaciones entre palabras y conceptos?
- Qué tipo de información se debe extraer de las categorías que sea representativa?
- Cómo representar la información extraida?