Detección y seguimiento de temas de actualidad

Sistemas de Recuperación de Información y Recomendación

Lidia Sánchez Mérida

Índice

Definición y conceptos

La detección y seguimiento de temas consiste en analizar información de varias fuentes para identificar la **aparición de nuevos tópicos o la reaparición de algunos ya existentes.**

Fuentes
(canales de noticias)

Historias
(artículos, secciones)

Eventos
(tiempo y lugar específico con sus hechos y consecuencias)

Nuevos eventos
(comienza el seguimiento)

Eventos existentes
(como antecedentes o consecuentes)

Origen de los sistemas TDT

- Gran cantidad de información multimedia (audio, texto...).
- Necesidad de organizar los datos de manera eficiente y automática para detectar eventos novedosos y realizar un seguimiento.
  - Navegadores web.
  - Buscadores.
  - Sistemas de alertas.
- Su aparición se ve favorecida por el desarrollo de las siguientes técnicas:
  - **Algoritmos de Aprendizaje Automático.**
  - **Técnicas para el Procesamiento del Lenguaje Natural.**
  - **Técnicas de Recuperación de Información.**

Componentes y técnicas: Segmentación

- Transforma un flujo de información en un conjunto de historias en dos etapas:
  - **Identificación de fronteras**: probabilidad de que una palabra o frase actúe como frontera entre dos historias en un vecindario.
  - **Refinamiento de los resultados**: agrupa historias similares a partir de su contenido.
- Se apoya en técnicas de procesamiento y análisis morfológicos para detectar el inicio y fin de las historias (silencios, signos de puntuación...).
- Evaluación de los resultados:
  - **Técnicas directas:** probabilidad de que una palabra o frase pertenezcan correctamente a la misma historia.
    - Conteo de falsos positivos (*extra boundary*) y negativos (*miss boundary)*.
  - **Técnicas indirectas:** se evalúa la calidad de la segmentación a partir del rendimiento asociado a la detección y seguimiento de nuevos eventos.
    - Grado de solapamiento entre historias.

Componentes y técnicas. Detección de eventos

- Identifica si las nuevas historias que aparecen en un flujo de datos contienen eventos desconocidos por el sistema.
- ***Retrospectiva***.
  - Identifica todos los eventos considerando todas las historias.
  - Cada evento es un cluster y cada historia debe pertenecer al menos a un cluster.
  - Evaluación: correspondencia entre historias~eventos y historias~clústeres.
- ***Online***.
  - Identifica los eventos utilizando el flujo de historias en tiempo real comprobando si contiene nuevos eventos.
  - Evaluación.
    - Menor número de muestras para testar los clústeres resultantes.
    - Modifican el corpus de historias dinámicamente y reservar una parte para test.

Componentes y técnicas. Seguimiento de eventos

- Asocia las nuevas historias que aparecen con eventos conocidos.
- Cada evento conocido dispone de una lista de historias relacionadas.
- Problema de clasificación a partir de los clústeres compuestos durante el proceso de detección.
  - Un clasificador binario por evento diferente.
    - Si la historia habla del evento o no.
    - El grado de confianza de que la historia habla sobre el evento.
  - Respeta el orden cronológico de las historias al validar el clasificador.
- ***Tracking as Detection***.
  - Similitud entre el sistema de detección y seguimiento de tópicos.
  - Objetivo: utilizar el sistema de detección como seguimiento también.
  - Una historia solo puede pertenecer a un único clúster de detección pero a varios de seguimiento.

Aplicaciones de los TDT

- Identificar tópicos y noticias a partir de flujos de datos.
    - Audio: resumen de temas a partir de transcripciones de audios.
    - Texto: organizar documentos por temáticas.
- *Web spiders.*
- Mercados financieros.
- Control de recursos en plataformas cloud.

Bibliografía

STORY SEGMENTATION AND TOPIC DETECTION FOR RECOGNIZED SPEECH, S. Dharanipragada, M. Franz, J.S. McCarley, S. Roukos, T. Ward

Topic Detection and Tracking Pilot Study Final Report, James Allan, Jaime Carbonell, George Doddington, Jonathan Yamron, Yiming Yang

The Design of a Topic Tracking System, Joe Carthy, Alan F. Smeaton

Emerging topic tracking system in WWW, Khoo Khyou Bun, Mitsuru Ishizuka

Online Topic Detection and Tracking System and Its Application on Stock Market in China, Yuefeng Lin, Zhongchen Miao, Mengjun Ni, Hang Jiang, Chenyu Wang, Jian Gao, Jidong Lu, Guangwei Shi

Emerging Topic Tracking System, Mitsuru Ishizuka

A multi-layered performance analysis for cloud-based topic detection and tracking in Big Data applications, Meisong Wang-Prem, Prakash Jayaraman, Ellis Solaiman, Lydia Y.Chen, ZhengLi, Song Jun, Dimitrios Georgakopoulos, Rajiv Ranjan

Gracias por la atención