Búsqueda y Recuperación de Información en Textos

Víctor Mijangos de la Cruz

VI. Aplicaciones de RI

Sistemas de recomendación

Necesidades de los usuarios

Los usuarios pueden tener varios tipos de necesidades de información:

- **Necesidades de corto plazo:** Responden a una necesidad de información inmediata, que puede solucionarse con una búsqueda dentro de un sistema de RI.
- **Necesidades a largo plazo:** Responden a necesidades continuas y/o que surgen a partir de uso prolongado del sistema. Por ejemplo, filtrar información.

Filtrado de documentos

Filtrado de documentos

En el filtrado de documentos, los documentos se obtienen de una fuente dinámica (correo electrónico, páginas web). Un sistema de filtrado toma una decisión binaria en base a la relevancia de un documento para un usuario en cuanto el documento es considerado.

La pregunta esencial del filtrado de documentos puede plantearse como:

¿Es el documento d de interés para el usuario u?

Algunos sistemas de recomendación comunes son:

- Filtrado de correos electrónicos con **spam**.
- Filtrado de documentos mal-intencionados.

Estrategias de filtrado

Consideramos dos estrategias de filtrado que consideran a los documentos d y el usuario u.

Filtrado basado en contenido: Determina cuál es el contenido que el usuario u prefiere y caracteriza a d con base en esto.

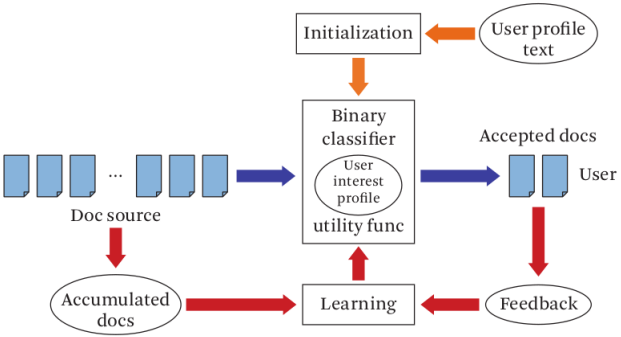
Filtrado colaborativo: Determina la relevancia de d en base a los gustos que otros usuarios tienen sobre d. Si hay correspondencia entre los usuarios que gustaron de d y u se recomienda d.

Filtrado basado en contenido

El filtrado basado en contenido se basa en los siguientes elementos:

- Una **fuente de los documentos:** La fuente de donde provienen o se generan los documentos.
- Un **perfil del usuario:** Determina los intereses del usuario.
- Una **función de utilidad:** que ayuda al sistema a tomar decisiones.
- Un **clasificador binario:** Que es el centro de las decisiones de filtrado.
- **Retroalimentación:** Retroalimentación que el sistema recibe del usuario para mejorar el filtrado y las recomendaciones.

Filtrado basado en contenido



Perfil de usuario

El perfil de usuario busca caracterizar a un usuario u según sus preferencias sobre la información.

Este perfil puede formarse a partir de la **actividad** que el usuario tiene con el sistema (los documentos que revisa, los que deshecha, etc.)

En principio, antes de interactuar con el sistema, el perfil puede basarse en:

- Un resumen general que caracterice al usuario.
- Una lista de palabras clave que pueda caracterizar los gustos del usuario.

Función de utilidad

Una **función de utilidad** determina un **umbral** θ de aceptación o bien, decide si el documento debe o no mostrarse al usuario.

Función de utilidad lineal

Sea R el conjunto de documentos recuperados que han sido relevantes para el usuario y R^c aquellos que no lo son, una función de utilidad se puede definir como la combinación lineal: $\mathcal{U} = 3 \cdot |R| - 2 \cdot |R^c|$

El objetivo es maximizar la función de utilidad; es decir, recomendar el mayor número de documentos que son realmente necesarios para el usuario, mientras que no recomendar aquellos que rechazará.

Clasificación y retroalimentación

La clasificación se puede hacer con un clasificador binario que determine si es relevante o no un documento.

Generalmente se utilizan **modelos de aprendizaje**, como Bayes ingenuo, SVMs, Árboles de decisión, redes neuronales, etc.

La **retroalimentación** consiste en tomar en cuenta las decisiones del usuario con respecto a los documentos recomendados para mejorar la clasificación.

Se realiza un **seguimiento** de la actividad del usuario para poder alimentar una base de datos de documentos que ayuden a que la clasificación sea mejor.

Módulos del filtrado basado en contenido

Trss módulos esenciales se dan en el filtrado basado en contenido:

- **Módulo de inicialización:** Inicia el sistema basado en únicamente una descripción del texto limitada, o de pocos ejemplos del usuario.
- **Módulo de decisión:** Dado un perfil de usuario, decide si el documento será recomendado o no.
- **Módulo de aprendizaje:** Aprende (mejora la decisión) en base a las decisiones que toma el usuario, y su interacción con el sistema.

Filtrado por score

Una forma común de realizar un primer filtrado (sobre todo en la inicialización) es a partir de un **filtrado por score**.

En base a un sistema de recuperación, se puede calcular un score $score(d_i)$ sobre el documento d_i, de tal forma que la decisión será:

$$f(d_i) = \begin{cases} 1 & \text{si } score(d_i) > \theta \\ 0 & \textbf{si } score(d_i) \leq \theta \end{cases}$$

En este caso el score se determina como:

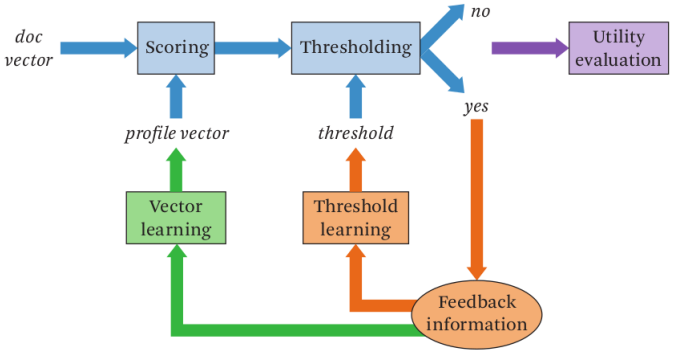
$$score(d_i) = \phi(\overrightarrow{d_i}, \overrightarrow{u})$$

Donde $\overrightarrow{d_i}$ es un vector del documento y $\overrightarrow{u}$ es un vector del perfil de usuario.

Maximización de la utilidad

Para maximizar la utilidad se tienen dos casos:

- Se puede mejorar la función de score para que se adapte al umbral.
- Se puede modificar el umbral para que se mejore la decisión con la función de score.



Exploración-explotación

Para ajustar el sistema y que podamos aprender es necesario **explorar**, pero explorar implicaría recomendar al usuario documentos no pertinentes.

Compensación de exploración-explotación

El problema de la compensación entre exploración y explotación consiste en buscar un umbral adecuado que permita explorar los gustos del usuario, pero sin recomendar un número excesivo de documentos no relevantes.

El objetivo es entonces poder explorar lo suficiente para aprender, pero sin agobiar al usuario con recomendaciones no relevantes.

Aprendizaje de umbral

Para aprender un umbral adecuado, este se puede adaptar en base al siguiente método ($\beta, \gamma \in [0, 1]$):

Algorithm Aprendizaje de umbral

procedure THRESHOLDLEARNING(β, γ, u)
  $\theta_{opt} \leftarrow \max_i\{$ UTILITY$(d_i; u)\}$
  $\theta_0 \leftarrow d_i$ tal que $score(d_i; u) = 0$
  $\alpha \leftarrow \beta + (1 - \beta) exp\{-\gamma N\}$
  $\theta^* \leftarrow \alpha\theta_0 + (1 - \alpha)\theta_{opt}$
  $u \leftarrow$ UPDATE$(u; \theta^*)$
  return θ^*
end procedure

Filtrado colaborativo

El **filtrado colaborativo** hace uso de las opiniones de otros usuarios para hacer recomendaciones a un usuario particular.

Se debe encontrar un conjunto de **usuarios similares**; que tengan un comportamiento parecido. Esto se basa en:

- Usuarios con intereses comunes, tienen preferencias similares.
- Usuarios con la misma preferencia, comparte el mismo interés.

Ranking de usuarios

Dado un usuario u, se clasifican otros usuarios $u_1, u_2, ..., u_m$ basados en la similitud con el usuario u.

Se consideran las **preferencias** de los usuarios: esto es, un conjunto de objetos $o_1, ..., o_n$.

Si existe un juicio del usuario u_i por el objeto o_j, se manifiesta en una calificación $R_{i,j}$.
Esta configuración define una matriz determinada como:

$$A = (a_{i,j}) = f(u_i, o_j) = R_{i,j}$$

La función f determina la relevancia del objeto o_j para el usuario u_i; puede ser una calificación, u otra función que determine numéricamente la relevancia.

Recomendación colaborativa

Sea μ_i la media de valoración para el usuario u_i; es decir:

$$\mu_i = \frac{1}{|n|} \sum_j R_{i,j}$$

Y sea $\nu_{i,j}$ el valor normalizado de las valoraciones del usuario u_i para el objeto o_j:

$$\nu_{i,j} = R_{i,j} - \mu_i$$

Entonces, el rango normalizado que se predice en la recomendación colaborativa para el usuario u es:

$$\hat{\rho}_{u,j} = \frac{1}{Z} \sum_{i=1}^{m} w(u, u_i) \cdot \nu_{i,j}$$

Donde $Z = \sum_i w(u, u_i)$ y la función w es una función de similitud entre usuarios.

Recomendación colaborativa

La predicción hecha puede transformarse en un rango para el usuario u:

$$\hat{R}_{u,j} = \hat{\rho}_{u,j} + \mu_u$$

La función de pesos w puede determinarse de varias formas, como por medio de los vectores de usuario, o partir de la matriz de usuarios y preferencias:

$$w(u_k, u_i) = \frac{\sum_j (R_{k,j} - \mu_k)(R_{i,j} - \mu_i)}{\sqrt{\sum_j (R_{k,j} - \mu_k)^2 \sum_j (R_{i,j} - \mu_i)^2}}$$

Minería de opiniones y análisis de sentimientos

Objetivo y subjetivo

Las declaraciones **objetivas** responden a circunstancias factuales del mundo (hechos) que pueden probarse verdaderos o falsos.

Un ejemplo de declaración subjetiva puede referir a hechos concretos del mundo:

La batería del celular se acabó.

Este tipo de declaraciones pueden ser comprobadas, de tal forma que se afirme su veracidad.

Las declaraciones **subjetivas** reflejan lo que una persona piensa/siente acerca de algo.

No pueden ser probadas verdaderas o falsas. No tienen valores de verdad. Por ejemplo:

Este celular es más elegante.

Estas declaraciones no pueden ser comprobadas en el mundo, y dependen de un **sujeto** que las sustente.

Opiniones

Dentro de los juicios subjetivos encontramos las opiniones.

Opinión

Una opinión es un juicio subjetivo que describe lo que una persona cree o piensa acerca de algo.

La opinión, entonces, está sustentada por los siguientes elementos:

- Una **persona** o alguien que sustenta la opinión.
- Un **objetivo** que es el algo sobre lo que se opina.
- La **creencia o pensamiento** que se basa en la cultura, contexto, conocimientos previos.

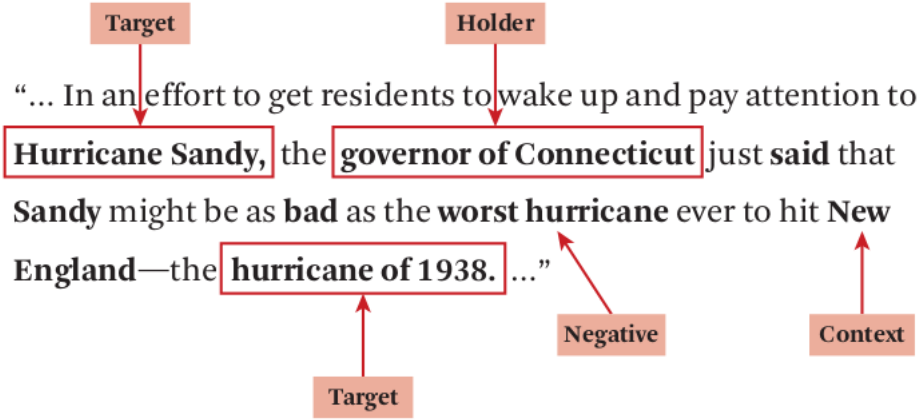
Clasificación de opiniones

Las opiniones conllevan un **sentimiento** que, en general, puede categorizarse en tres tipos:

- **Positiva:** Declara que el sustentante de la opinión tiene sentimientos positivos acerca del objetivo.
- **Negativa:** Declara que el sustentante de la opinión tiene sentimientos negativos acerca del objetivo.
- **Neutra:** Refiere a que no se determina un sentimiento, y por tanto, no se suele considerar una opinión.

Estos sentimientos no son necesariamente discretos. Pueden usarse **escalas continuas** que vayan de lo negativo a lo positivo.

Componentes de la opinión



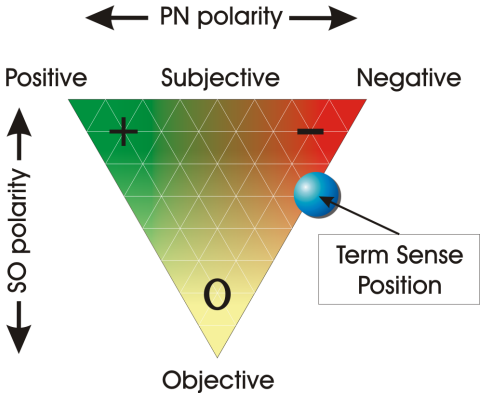
Clasificación de opinión

El objetivo de la clasificación de opinión es clasificar un texto dentro de una de las categorías positivo, negativo o neutro.

Generalmente, se asume que la opinión depende de las palabras que se usan en el texto (y cómo se usan).

- **Palabras con opinión positiva:** Bueno, mejor, excelente, agradable, etc.
- **Palabras con opinión negativa:** Malo, peor, pésimo, desagradable, etc.

Polaridad


PN polarity
Positive
Subjective
Negative
SO polarity
+
−
O
Term Sense Position
Objective

Herramientas para clasificación de opinión

Algunas de las herramientas disponibles para realizar la clasificación de opinión y análisis de sentimiento son:

- **SentiWordNet:** Lexicón para análisis de sentimiento que caracteriza a una palabra em tres ejes: positivo, negativo, objetivo. https://github.com/aesuli/SentiWordNet.
- **MonkeyLearn:** Sistema para análisis de sentimientos personalizable. https://monkeylearn.com/sentiment-analysis/.
- **Datasets:**
  - Dataset de reseñas de películas de IMDB https://datasets.imdbws.com/
  - Dataset de productos de Amazon: http://jmcauley.ucsd.edu/data/amazon/
  - Sentiment 140 http://help.sentiment140.com/for-students

Lidiando con texto espontáneo

La clasificación de opiniones y análisis de sentimiento generalmente proviene de **fuentes espontáneas** (mensajería instantánea, redes sociales).

Los textos espontáneos suelen tener características que los distinguen de textos no espontáneos:

- Faltas de ortografía o errores tipográficos.
- Alargamientos de sílabas como 'holaaaaaa' que pueden expresar emociones.
- Emoticones usados para expresar alguna emoción :(ó :)

Manejo de texto espontáneo

Para manejar el texto espontáneo es recomendado trabajar con **n-gramas de carácter**; es decir, tomar n caracteres en lugar de palabras.

Por ejemplo del texto:

$$\textit{Me gusta :)}$$

Se obtienen los 3-gramas:

$$\{\textit{me}, \textit{gus}, \textit{ust}, \textit{sta}, :)\}$$

Esto ayuda a lidiar con formas particulares. También se puede usar un vocabulario y usar alguna métrica entre cadenas (como Levenshtein).

Métodos de clasificación

La clasificación de sentimiento/opinión puede verse como una clasificación binaria (positivo, negativo) o terciaria (positivo, negativo, neutro):

- En el caso **binario** se puede selecciona un clasificado (regresión logística, Bayes ingenuo, SVMs, redes neuronales, etc.).
- En el caso **terciario** puede: 1) hacer una clasificación en tres clases con cualquier clasificador; o 2) Usar un clasificador binario que distinga entre **objetivo-subjetivo** y sólo los elementos subjetivos se clasificarán con otro clasificador binario **positivo-negativo**.

Análisis de aspectos latentes

Análisis de aspectos latentes

El análisis de aspectos latentes busca, en base a un conjunto de reseñas, busca encontrar aspectos latentes dentro de las reseñas generando calificaciones para estas, así como asignarle un peso (relevancia) a los aspectos.

Este tipo de análisis busca determinar tres cosas:

- Los aspectos latentes relevantes a las reseñas.
- Las calificaciones que les corresponden a cada uno de los aspectos según la reseña.
- Los pesos de los aspectos, relativos al usuario.

Análisis de aspectos latentes

Hotel XXX

Reviewer 1: ★★★★★

"Great location + spacious room = happy traveler"

Stayed for a weekend in July. Walked everywhere, enjoyed comfy bed and quiet hallways....

Reviewer 2: ★★★★★

"Terrific service and gorgeous facility"

I stayed at the hotel with my young daughter for three nights June 17–20, 2010 and absolutely loved the hotel. The room was one of the nicest I've ever stayed in ...

How to infer aspect ratings?

How to infer aspect weights?

Análisis de aspectos latentes

El análisis se puede dividir en los siguientes pasos:

- **Segmentar** las reseñas en los posibles **aspectos**.
- Determinar los valores de **calificación** en base a un método de regresión.
- Estimar las **calificaciones** de cada uno de los aspectos en el documento.
- Inferir el **rating final** en base a las calificaciones de los aspectos.

Calificación de aspectos

Sea w una palabra que aparece en el aspecto i, podemos contar con un **lexicón** que asigne un peso de sentimiento de la palabra w en el aspecto i, $\beta_{i,w}$.

Además, se cuenta con un **peso de los aspectos**, el peso $\alpha_i(d)$ determina la relevancia del aspecto i en el documento d.

Asimismo, tenemos las **calificaciones de cada aspecto** i en el documento d que denotamos como $r_i(d)$. Podemos determinar este rating como:

$$r_i(d) = \sum_w c_i(w, d)\beta_{i,w}$$

Donde $c_i(w, d)$ es la frecuencia de la palabra w en el aspecto i del documento d.

Calificación general

El conjunto de documentos con el que contamos será supervisado, pues contará con un rango general r_d; el conjunto será:

$$\mathcal{S} = \{(d, r_d)\}$$

Donde d es un documento/reseña y r_d su rating general. En general, asumiremos que:

$$r_d \sim \mathcal{N}(\sum_i \alpha_i(d) r_i(d), \sigma^2)$$

Es decir, se vuelve un problema de regresión en donde el valor esperado del rating es una combinación lineal:

$$\sum_i \alpha_i(d) r_i(d)$$

Lecturas recomendadas

ChengXiang, Z. y Massung, S. (2014). "11. Recomendation Systems". *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Mining*. pp. 221-238.

ChengXiang, Z. Jansen, p., Stoica, E., Grot, N. y Evans, D. (1999). "Threshold Calibration in CLARIT Adaptive Filtering". *Zhai, Chengxiang, et al. "Threshold calibration in CLARIT adaptive filtering." In Proceeding of Seventh Text REtrieval Conference (TREC-7)*.

ChengXiang, Z. y Massung, S. (2014). "18. Opinion Mining and Sentiment Analysis". *Text Data Management and Analysis: A Practical Introduction to Information Retrieval and Text Minning*. pp. 289-412.

Liu, B. (2012). *Sentiment Analysis and Opinion Mining*. Claypool Publishers.