Desarrollo de una medida de similaridad para Sistemas de Recomendación en sitios de Community Question Answering. Análisis desde un enfoque Big Data y usando un método de ensamble de clustering

Ing. Federico Tesone

Universidad Tecnológica Nacional (FRR)

Fecha pendiente

Aca va el jurado y eso

Tabla de contenidos I

Tabla de contenidos II

Área temática

Este trabajo se basa en 5 pilares teóricos:

- Sitios de Community Question Answering (CQA).

Área temática

Este trabajo se basa en 5 pilares teóricos:

- Sitios de Community Question Answering (CQA).
- Medidas de similaridad.

Área temática

Este trabajo se basa en 5 pilares teóricos:

- Sitios de Community Question Answering (CQA).
- Medidas de similaridad.
- Sistemas de Recomendación.

Área temática

Este trabajo se basa en 5 pilares teóricos:

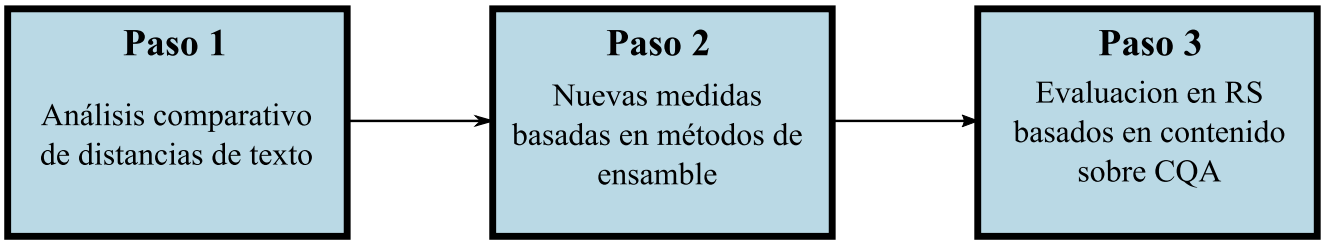
- Sitios de Community Question Answering (CQA).
- Medidas de similaridad.
- Sistemas de Recomendación.
- Big Data.

Área temática

Este trabajo se basa en 5 pilares teóricos:

- Sitios de Community Question Answering (CQA).
- Medidas de similaridad.
- Sistemas de Recomendación.
- Big Data.
- Ensamble de Clustering.

Tema específico I



Tema específico II

Considerando el conjunto completo de datos Quora (404301 pares de preguntas, es decir, 808602 preguntas totales), deberiamos realizar:

$\frac{n(n+1)}{2} = 326919001503$ calculos de distancias, donde $n = 808602$

Objetivo general

Construir una arquitectura Big Data que incluye la posibilidad de ser aplicada a grandes conjuntos de datos de preguntas en el ámbito de CQA y, a partir de esta arquitectura, implementar y evaluar nuevas medidas de similaridad entre textos que puedan ser utilizadas en sistemas de recomendación.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.
- Proponer una nueva medida que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data y que sea extensible a otras medidas existentes en el estado del arte.

Objetivos específicos

- Diseñar y desarrollar una arquitectura Big Data para cálculo de similaridad en grandes matrices, que requerirá nuevas estrategias para recolectar, procesar y manejar grandes volúmenes de datos.
- Identificar medidas de similaridad de texto existentes y un método efectivo de aplicación de las mismas en grandes volúmenes de datos.
- Evaluar el comportamiento de medidas de similaridad de texto del estado del arte respecto al manejo del volumen, variedad, velocidad y veracidad inherentes a grandes volúmenes de datos, en particular en el ámbito de CQA.
- Proponer una nueva medida que permita integrar las medidas de similaridad del estado del arte mediante una arquitectura de software basada en Big Data y que sea extensible a otras medidas existentes en el estado del arte.
- Brindar conclusiones, pautas y recomendaciones para trabajar con medidas de comparación de textos en grandes volúmenes de datos en sitios de CQA utilizando arquitecturas basadas en Big Data.

Motivación de la tesis

- Las medidas de similaridad del estado del arte tienen conocidos problemas.

Motivación de la tesis

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que puede además aplicarse como entrada para un RS con el fin de ser implementado en sitios de CQA.

Motivación de la tesis

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que puede además aplicarse como entrada para un RS con el fin de ser implementado en sitios de CQA.
- Es difícil identificar un algoritmo de Clustering que pueda manejar todos los tipos de formas y tamaños de cluster.

Motivación de la tesis

- Las medidas de similaridad del estado del arte tienen conocidos problemas.
- Creación de un método novedoso que combine medidas de similaridad existentes, que puede además aplicarse como entrada para un RS con el fin de ser implementado en sitios de CQA.
- Es difícil identificar un algoritmo de Clustering que pueda manejar todos los tipos de formas y tamaños de cluster.
- Arquitectura de software que soporte el procesamiento del método propuesto de una forma eficiente y escalable.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.
- Explorar desafíos tecnológicos que sirvan como fuente de futuras investigaciones o desarrollos de software.

Importancia científico-tecnológica

- Mejorar recomendaciones en sitios de CQA.
- Explorar desafíos tecnológicos que sirvan como fuente de futuras investigaciones o desarrollos de software.
- Aplicación de los conocimientos obtenidos en este trabajo para mejorar otros sitios basadas en busquedas textuales.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.
- Presentación en congresos tales como AGRANDA, CONAIISI, o RecSys.

Formación de recursos humanos

- Capacitar a grupos de estudiantes de la UTN FRRo.
- Presentación en congresos tales como AGRANDA, CONAIISI, o RecSys.
- Elaborar material de estudio relacionado con la temática para materias de grado, cursos, o disertaciones.

Sitios de CQA

Los servicios de *Community Question Answering* CQA, son un tipo especial de servicios de *Question Answering* (QA), los cuales permiten a los usuarios registrados responder a preguntas formuladas por otras personas.

Sitios de CQA

El mecanismo existente por el cual se responden las preguntas en los sitios de CQA todavía no alcanza a satisfacer las expectativas de los usuarios por varias razones:

- Baja probabilidad de encontrar al experto.

Sitios de CQA

El mecanismo existente por el cual se responden las preguntas en los sitios de CQA todavía no alcanza a satisfacer las expectativas de los usuarios por varias razones:

- Baja probabilidad de encontrar al experto.
- Respuestas de baja calidad.

Sitios de CQA

El mecanismo existente por el cual se responden las preguntas en los sitios de CQA todavía no alcanza a satisfacer las expectativas de los usuarios por varias razones:

- Baja probabilidad de encontrar al experto.
- Respuestas de baja calidad.
- Preguntas archivadas y poco consultadas: muchas preguntas de los usuarios son similares.

Sistemas de recomendación

Como se mencionó anteriormente, un RS es un conjunto de herramientas de software que sugiere ítems a un usuario, quien posiblemente utilizará algunos de ellos.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- Basados en contenido.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- Basados en contenido.
- Demográficos.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- Basados en contenido.
- Demográficos.
- Basados en conocimiento.

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- Basados en contenido.
- Demográficos.
- Basados en conocimiento.
- Basados en comunidades (sociales).

Técnicas de Recomendación

Los RS basan sus estrategias de recomendaciones en 6 técnicas básicas (Ricci et al., 2011):

- Basados en contenido.
- Basados en contenido.
- Demográficos.
- Basados en conocimiento.
- Basados en comunidades (sociales).
- Sistemas Híbridos.

Big Data

Conjuntos de datos cuyo tamaño está más allá de la habilidad de las herramientas software de base de datos para capturar, almacenar, gestionar y analizar los datos (Manyika et al., 2011).

"Big Data son activos de información caracterizados por su alto volumen, velocidad y variedad que demandan formas innovadoras y rentables de procesamiento de información para mejorar la compresión y la toma de decisiones" (consultora Gartner).

Big Data

Conceptos importantes relativos a este trabajo:

- Map-reduce.

Big Data

Conceptos importantes relativos a este trabajo:

- Map-reduce.
- Arquitectura Hadoop (HDFS).

Big Data

Conceptos importantes relativos a este trabajo:

- Map-reduce.
- Arquitectura Hadoop (HDFS).
- Apache Spark.

Information retrieval

Information retrieval (IR), traducido a menudo como "recuperación de información", se define la acción de como encontrar material (generalmente documentos) de una naturaleza desestructurada (generalmente texto) que satisfaga una necesidad de información de grandes colecciones (generalmente almacenadas en computadoras) (Schütze et al. 2008).

Similaridad

Las medidas de similaridad son interés poder cuantificar la relación entre objetos.

La función de similaridad es definida satisfaciendo las condiciones:

1. Simetría,
$$S(x_i, x_j) = S(x_j, x_i);$$
2. Positividad,
$$0 \leq S(x_i, x_j) \leq 1, \quad \forall x_i, x_j.$$

Es posible transformar una medida de similaridad $S(x_i, x_j)$ en una de distancia $D(xi, xj)$ que cumpla $0 \leq D(x_i, x_j) \leq 1$, en el intervalo $[0, 1]$. Aplicando $D(x_i, x_j) = 1 - S(x_i, x_j)$.

Modelo de espacio vectorial

En el modelo de *espacio vectorial*, un texto es representado como un vector de términos. Si las palabras son elegidas como términos, entonces cada palabra del vocabulario sería una *dimensión* independiente en el espacio vectorial (Singhal et al., 2001).

Típicamente, el ángulo entre los dos vectores es usado como medida de divergencia entre los mismos, y el coseno del ángulo es usado como similaridad numérica.

Distancia del coseno

Siendo $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$ dos documentos en forma de vectores:

$$d_c(\vec{D}_i, \vec{D}_j) = 1 - cos(\vec{D}_i, \vec{D}_j) = 1 - \frac{\vec{D}_i' \vec{D}_j}{\sqrt{\vec{D}_i' \vec{D}_i}\sqrt{\vec{D}_j' \vec{D}_j}}.$$

Particularmente en IR, es de interés el intervalo $[0, 1]$, entonces la distancia del coseno se puede derivar de la fórmula del *producto escalar*:

$$\vec{D}_i . \vec{D}_j = \left\| \vec{D}_i \right\| . \left\| \vec{D}_j \right\| .cos(\theta),$$

siendo $\left\| \overrightarrow{D_i} \right\|$ y $\left\| \overrightarrow{D_j} \right\|$ los módulos de los vectores $\overrightarrow{D_i}$ y $\overrightarrow{D_j}$ respectivamente, y θ el ángulo formado entre ellos

Term Frequency (TF)

- También conocido en la literatura como *Bag of words* (bolsa de palabras).
- El orden exacto de los términos es ignorado, pero se basa en el número de ocurrencias de cada uno de ellos en un documento.
- Cada documento corresponde a un vector y cada término a una dimensión.
- Se mide el grado de similaridad de dos documentos utilizando el coseno del ángulo.

"Mary is quicker than John" y *"John is quicker than Mary"*

Term Frequency (TF)

- Se define *document frequency* df_t como el número de documentos en una colección que contienen el término t.
- *Inverse document frequency*, o IDF, es un indicador basado en la cantidad de documentos que contienen (o son indexados por) un término en cuestión.
- Intuición: si un término de búsqueda se encuentra en muchos documentos, no es un buen discriminador, y se le debe asignar menor peso que a un término que se encuentra en pocos documentos.

$$tfidf(t_i, d_j) = tf(t_i, d_j) \cdot idf(t_j)$$

Word2Vec

- Modelos basados en redes neuronales con una capa oculta para computar representaciones de palabras como vectores continuos en grandes conjuntos de datos.
- Dos modelos: Skip-gram y Continuous Bag of Words.
- Las entradas y salidas de la red neuronal son palabras representadas como *one-hot* vector.
- Los pesos de la capa oculta se van ajustando utilizando un clasificador de regresión Softmax.
- Estos pesos resultantes dan como resultado a la representación vectorial de palabras utilizadas para el cálculo de similaridad de este trabajo.

FastText

Hipótesis I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

El método propuesto I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Arquitectura de procesamiento de datos I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Implementación en un sistema de recomendación de tiempo real I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Estado del arte I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Preprocesamiento y muestreo del conjunto de datos I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Preprocesamiento y muestreo del conjunto de datos I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Ensamble de Clustering I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Método de validación I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Análisis del método propuesto I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Análisis del método propuesto y algoritmos del estado del arte I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Análisis del método propuesto y algoritmos del estado del arte I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Análisis de desempeño I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Análisis de desempeño I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Contribuciones realizadas I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

Futuras investigaciones I

Lorem ipsum dolor sit amet, consectetur adipisicing elit, sed do eiusmod tempor incididunt ut labore et dolore magna aliqua.

References

John Smith (2012)

Title of the publication

Journal Name 12(3), 45 – 678.

The End