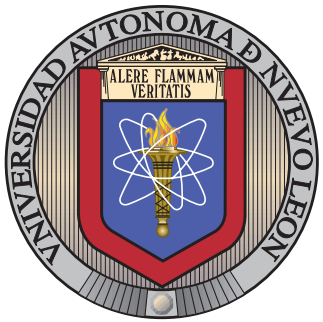
Reconocimiento de Entidades para Filtrado de Duplicados

Jorge Alberto Cordero Cruz

Universidad Autónoma de Nuevo León



21 de enero del 2014

Agenda

- Motivación para realizar detección de documentos duplicados
- Hipótesis
- Objetivos
- Reportes ciudadanos
- Sistema de detección de duplicados
- Resultados
- Trabajos relacionados
- Estado actual de la tesis

Motivación para realizar detección de documentos duplicados

Hoy en día existe cierto nivel de presión para que los investigadores de las universidades produzcan artículos de investigación. En ocasiones los investigadores plagian trabajos de otras personas. Por lo que una herramienta que permita detectar artículos científicos duplicados ayudaría a bajar los niveles de plagio.

Todos los días el Centro de Integración Ciudadana (CIC) recibe reportes (reportes ciudadanos) de problemas de vialidad y tránsito, situaciones de riesgo, accidentes, etc. Varias personas pueden reportar un mismo suceso por lo que los reportes recibidos pueden ser duplicados. Una herramienta capaz de detectar reportes duplicados permitiría agilizar el proceso de respuesta del CIC ante los reportes.

Hipótesis

- Las técnicas de etiquetado y reconocimiento de entidades pueden ser utilizadas para generar una representación estructurada a partir del contenido de diferentes tipos de documentos.
- Una representación estructurada permite utilizar técnicas de minería de datos y aprendizaje máquina sin tomar en cuenta los tipos de documentos con los que se trabaja.
- Las técnicas de minería de datos y aprendizaje máquina pueden ser utilizadas para detectar documentos duplicados en un repositorio.

Objetivos

El objetivo general consiste en detectar de manera automática documentos duplicados en un repositorio de documentos.

Objetivos específicos:

- Extraer de diferentes tipos de documentos la información referente a:
  - tiempos
  - lugares
  - información importante
  - nombres de personas y nombres de organizaciones
- Crear una representaciones estructuradas partir de la información extraida.
- Detectar documentos duplicados aplicando algoritmos supervisados a las representaciones estructuradas.

Reportes ciudadanos del CIC

En el trabajo desarrollado se utilizaron los reportes del CIC, los cuales fueron descargados desde la plataforma para desarrolladores CICMty-API en formato JSON. En la Tabla 1 se muestran las diferentes categorías a las que pertenecen los reportes ciudadanos.

Grupos	**Categorías**
Comunidad	Avisos, Evento público, Observador ciudadano
Emergencia	Emergencias
Propuestas ciudadanas	Propuesta comunidad, Propuesta seguridad, Propuesta servicios públicos, Propuesta vialidad
Seguridad	Incendio, Robo, Situación de riesgo
Servicios públicos	Alcantarillas, Alumbrado público, Falta electricidad, Fuga, Otros, Parques descuidados, Recolección de basura
Vialidad y tránsito	Accidente, Bache o vía dañada, Obras y/o vía cerrada, Semáforo descompuesto, Vialidad

Tabla 1: Grupos y categorías de grupo a las que puede pertenecer un reporte ciudadano

Categorías de reportes seleccionadas

Para este trabajo de tesis se seleccionaron reportes pertenecientes a las siguientes categorías:

- `Accidente`
- `Alumbrado público`
- `Semáforo descompuesto`
- `Bache o vía dañada`
- `Evento público`
- `Alcantarillas`
- `Obras y/o vías cerradas`
- `Situación de riesgo`

Ejemplo de reportes duplicados

Primer reporte

ACCIDENTE Unidad de la S.V.T. de Monterrey acude al lugar.
En Lincoln y Aguaturma Paseo de Las Mitras 64118 Monterrey NL Mexico.
2013-06-20 09:59:28

Segundo reporte

ACCIDENTE transito de Monterrey acude.
En Aguaturma y Lincoln Paseo de Las Mitras Monterrey Nuevo Leon Mexico.
2013-06-20 09:49:46

Ejemplo de reportes no duplicados

Primer reporte

ACCIDENTE Unidad de la S.V.T. de Monterrey acude al lugar.

En Lincoln y Aguaturma Paseo de Las Mitras 64118 Monterrey NL Mexico.

2013-06-20 09:59:28

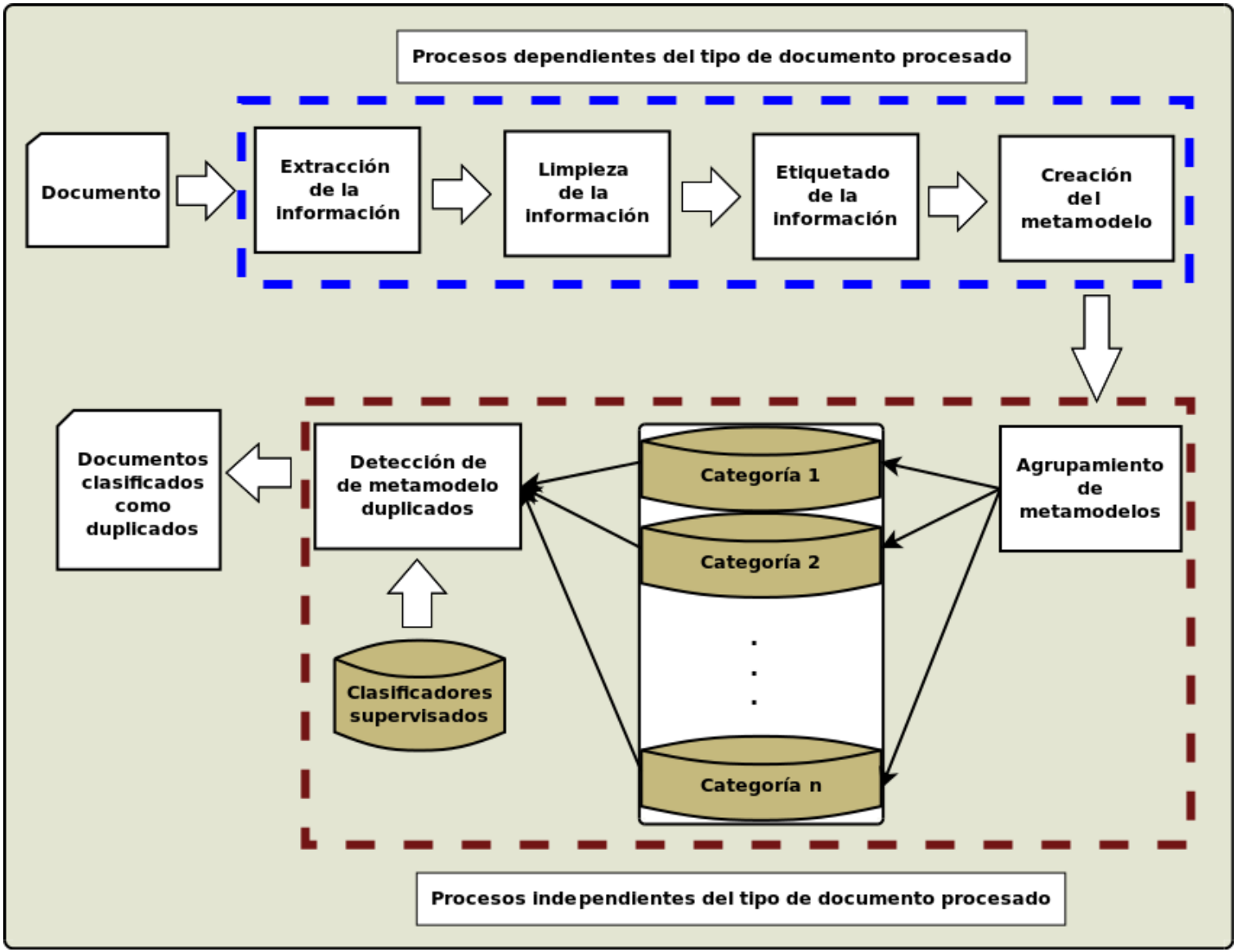
Segundo reporte

ACCIDENTE choque de motos.

Avenida Lincoln y Aguaturma P. de Las Mitras Monterrey NL Mexico.

2013-06-20 15:59:28

Sistema propuesto



Reporte ciudadano en formato JSON

```
{
    "ticket":"#7YPC",
    "content":"*ACCIDENTE* En gonzalitos altura de vuelta izquierda
               a Insurgentes MTY #mtyfollow 17:37",
    "created_at":"2013-08-04T17:49:56-05:00",
    "address_detail":{
        "formated_address":"Gonzalitos 655, Sin Nombre de Colonia 31,
                            Monterrey, NL, México",
        "zipcode":"64000",
        "county":{
            "long_name":"Monterrey",
            "short_name":"Monterrey",
        },
        "state":{
            "long_name":"Nuevo León",
            "short_name":"Nuevo León"
        },
        "neighborhood":{
            "long_name":"Centro",
            "short_name":"Centro"
        }
    },
    "group":"Vialidad y Transito (SS)",
    "categories":["ACCIDENTE"]
}
```

Extracción de la información

Los campos utilizados para extraer la información de los reportes son los siguientes:

- `'ticket'` (número de ticket).
- `'content'` (contenido).
- `'created_at'` (fecha de creación).
- `'address_detail':'formatted_address'` (dirección con formato).
- `'categories'` (categorías).

Limpieza de la información

Al texto extraído se aplican filtros para eliminar el ruido. Algunos de los filtros aplicados son los siguientes:

- Eliminación de acentos
  - `público` → `publico`
- Sustitución del punto (.) y dos puntos (:)
  - `ave.` → `ave_dot_`
  - `P. Livas` → `P_dot_ Livas`
  - `11:16` → `11_dot_dot_16`
- Eliminación de símbolos
  - ‘‘@’’
  - ‘‘_’’
  - ‘‘*’’

Limpieza de la información

- Eliminación de palabras con expresiones regulares, por ejemplo hashtags:
  - `r'#[a-zA-Z_0-9]+'` → `#1Accidente`
- Separación de oraciones usando el punto final (.), por ejemplo: ``Choque en avenida Garza Sada. Enviar paramédicos.''
  - Choque en avenida Garza Sada
  - Enviar paramédicos
- Sustitución de _dot_ y _dot_dot_
  - `ave_dot_` → `ave.`
  - `P_dot_ Livas` → `P. Livas`
  - `11_dot_dot_16` → `11:16`

Ejemplo de texto original y texto limpio

Texto original

ACCIDENTE en Ave. Garza Sada sin lesionados paramédicos acuden, 6:30 pm MTY NL #choquefuerte @cicmty via @colli03 gracias

Texto limpio

ACCIDENTE en Ave. Garza Sada sin lesionados paramedicos acuden, 6:30 pm MTY NL gracias

Etiquetado de la información

El contenido de los textos limpios se etiqueta para identificar nombres de entidades correspondientes a `lugares`, `tiempo` e `información relevante`. En este trabajo de tesis se uso un etiquetador muy sencillo basado en un Modelo Oculto de Markov.

Las etiquetas utilizadas en esta tesis fueron:

- `LOC` para indicar un lugar
- `TIME` para indicar tiempo
- `NAME` para indicar nombres de personas
- `ORG` para indicar nombres de organizaciones y/o empresas
- `TTERM` para indicar información de un suceso reportado
- `O` para indicar información irrelevante

Ejemplo de reconocimiento de entidades

Del texto limpio ‘‘ACCIDENTE en Ave. Garza Sada sin lesionados paramedicos acuden, 6:30 pm MTY NL gracias’’ se obtienen las siguientes entidades:

- LOC: en, Ave., Garza, Sada y NL
- TIME: 6:30 y pm
- TTERM: ACCIDENTE, sin, lesionados y paramedicos
- ORG: gracias

Creación del metamodelo

Un metamodelo es una representación estructurada del contenido de un texto etiquetado.
El metamodelo permite aplicar los mismos algoritmos de detección de duplicados a documentos de diferentes formatos (documentos PDF, páginas web, documentos JSON).

Recuadro 1: Metamodelo generado a partir de un texto etiquetado

```
<?xml version="1.0" encoding="ISO-8859-1">
<metamodelo>
    <tterm>ACCIDENTE sin lesionados paramedicos</tterm>
    <loc>en Ave. Garza Sada</loc>
    <time>6:30 pm</time>
</metamodelo>
```

Creación del metamodelo

Para cada tipo de documento que se procesa en el sistema se debe implementar el proceso de obtención del metamodelo mostrado en la Figura 2.

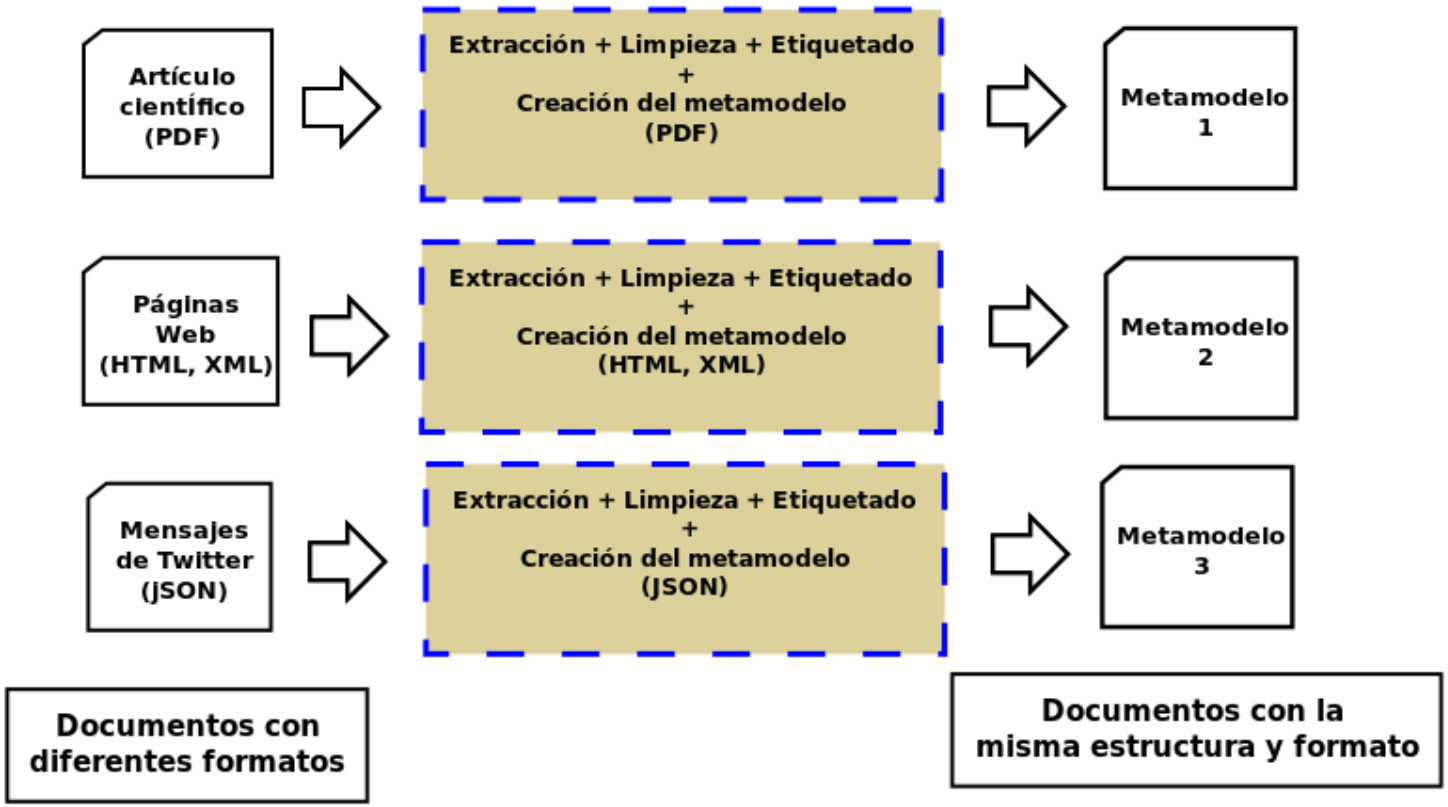


Figura 1: Generación de metamodelos para documentos de diferentes formatos

Agrupamiento de metamodelos

En el agrupamiento se juntan los metamodelos que son muy similares entre ellos; la Figura 3 ilustra esto.

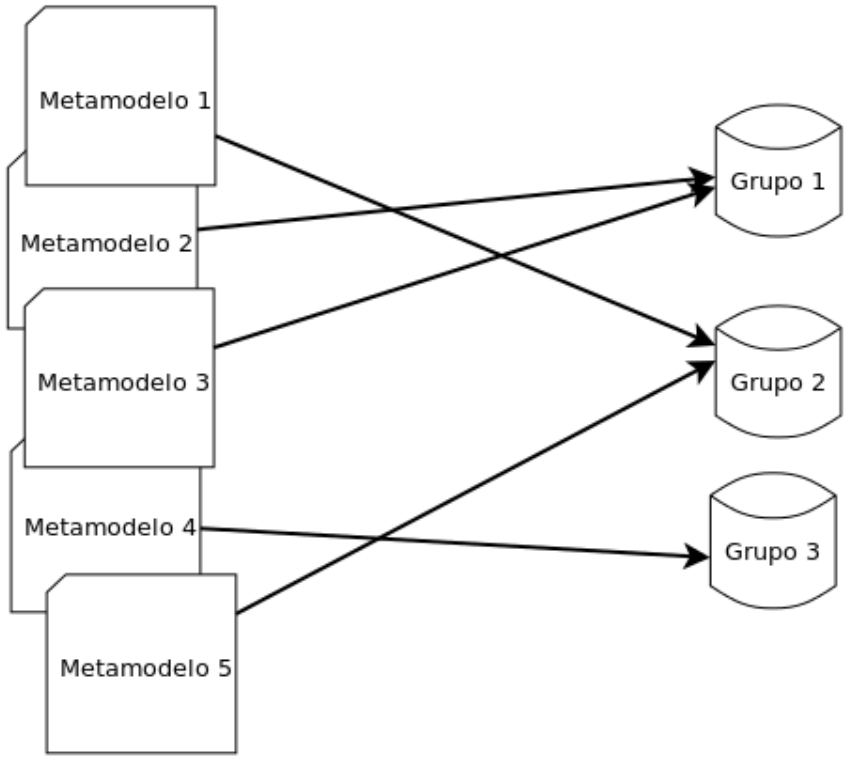


Figura 2: Agrupamiento de metamodelos

Agrupamiento de metamodelos

Se manejan dos maneras de realizar el agrupamiento:

- Si los metamodelos tienen asignada una de las ocho categorías de reporte seleccionadas, se agrupan por categoría.
- Cuando no se tiene la categoría del metamodelo se realiza agrupamiento no supervisado utilizando el algoritmo k-medias.

Entrenamiento de clasificadores

Para cada uno de los grupos de metamodelos se entrena un clasificador supervisado de tipo Máquina de soporte vectorial (SVM por sus siglas en inglés) de la siguiente manera:

- A partir de metamodelos de entrenamiento se obtiene una lista de tripletas

$$le = [(m_1^1, m_2^1, dup^1), (m_1^2, m_2^2, dup^2), \ldots, (m_1^{ne}, m_2^{ne}, dup^{ne})], \quad (1)$$

  donde m_1^i y m_2^i son metamodelos diferentes y $dup^i \in [0, 1]$ indica con 1 duplicado y con 0 no duplicado.

- Se calculan:
  - similitud de información *sinfo*
  - similitud de lugar *slugar*
  - diferencia de tiempo *dtiempo*

Entrenamiento de clasificadores

- Se crea la matriz de características X y el vector de etiquetas $\vec{y}$

$$X = \begin{bmatrix} sinfo^1, slugar^1, dtiempo^1 \\ sinfo^2, slugar^2, dtiempo^2 \\ \vdots \\ sinfo^{ne}, slugar^{ne}, dtiempo^{ne} \end{bmatrix} \tag{2}$$

$$\vec{y} = \begin{bmatrix} dup^1 \\ dup^2 \\ \vdots \\ dup^{ne} \end{bmatrix} \tag{3}$$

- Finalmente se entrena el clasificador

Detección de duplicados

Los metamodelos que son agrupados se comparan contra cada uno de los metamodelos del grupo utilizando el clasificador correspondiente.

Si el clasificador dicta que dos metamodelos son duplicados, entonces los reportes correspondientes son considerados duplicados. La Figura 4 ilustra la detección de duplicados para dos metamodelos.

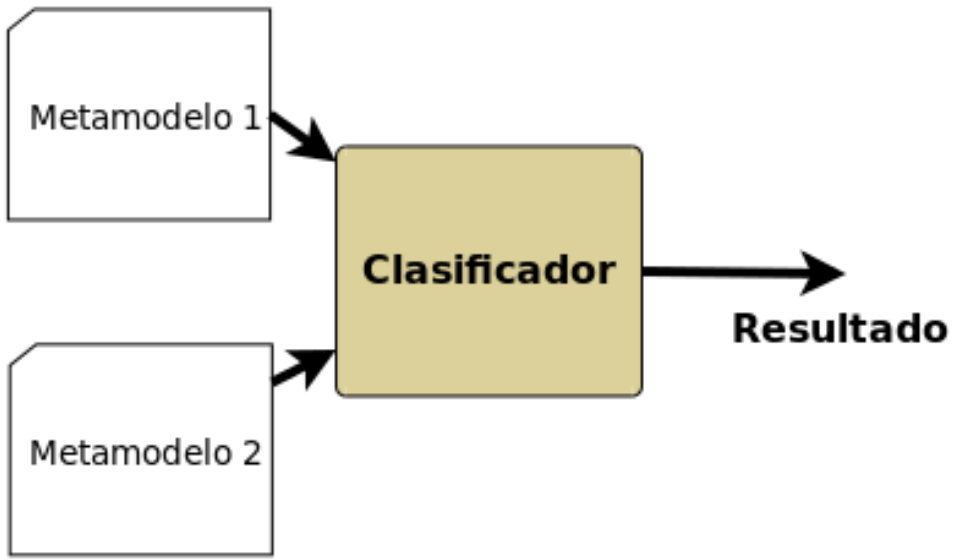


Figura 3: Detección de metamodelos duplicados

Resultados del etiquetador

Para probar el desempeño del etiquetador se utilizaron dos archivos de 5099 palabras cada uno.

El contenido de un archivo fue etiquetado manualmente y para el otro archivo se utilizo el etiquetador.

Se compararon las etiquetas obtenidas por el etiquetador contra las etiquetas asignadas manualmente y se obtuvo una precisión del 92%.

Para documentos obtenidos de Twitter esta precisión es muy buena y como los reportes son parecidos a los tweets, se considera un buen nivel de precisión obtenido por el etiquetador.

Resultados de la detección de duplicados

Se probó el desempeño de la detección de duplicados para dos enfoques diferentes:

- Enfoque no supervisado o híbrido; cuando no se conocen las categorías de los metamodelos durante el agrupamiento.
- Enfoque supervisado; cuando se conocen las categorías de los metamodelos durante el agrupamiento.

Resultados de la detección de duplicados

Para el enfoque híbrido se obtuvo un valor promedio de precisión del 46.9% y para el enfoque supervisado del 54.5%.

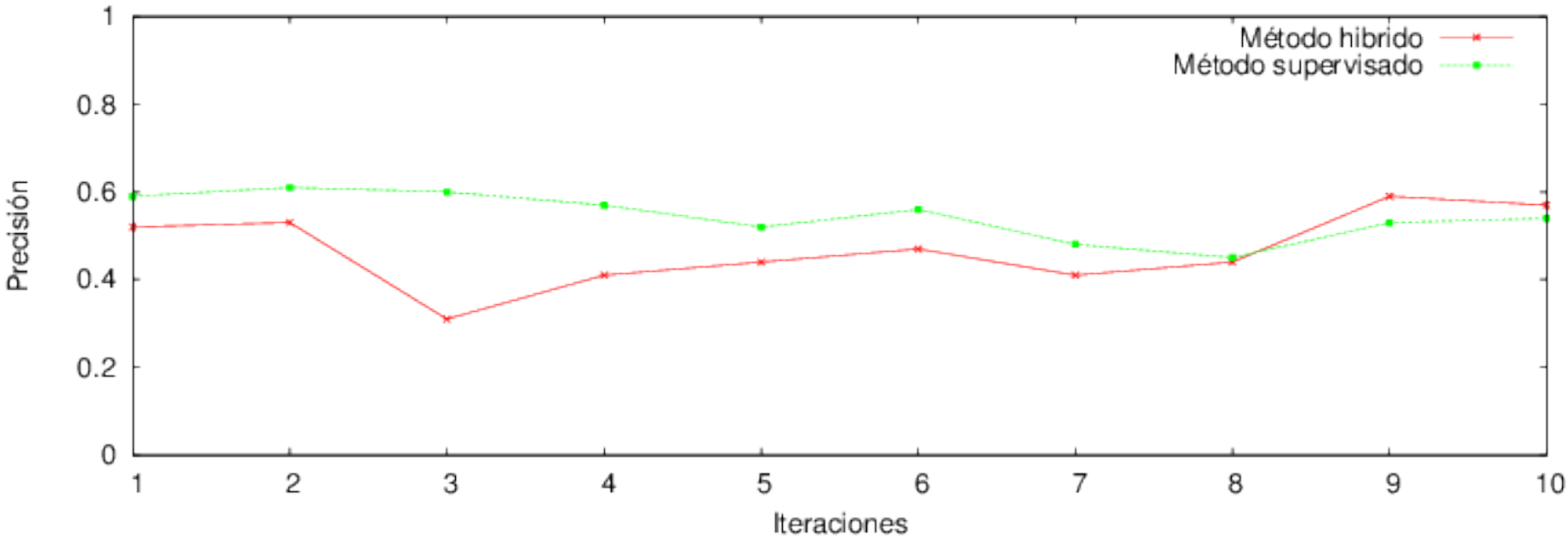


Figura 4: Valores de precisión obtenidos en la detección de duplicados

Resultados de la detección de duplicados

Para el enfoque híbrido se obtuvo un valor promedio del valor-F del 59.6% y para el enfoque supervisado del 65.6%.

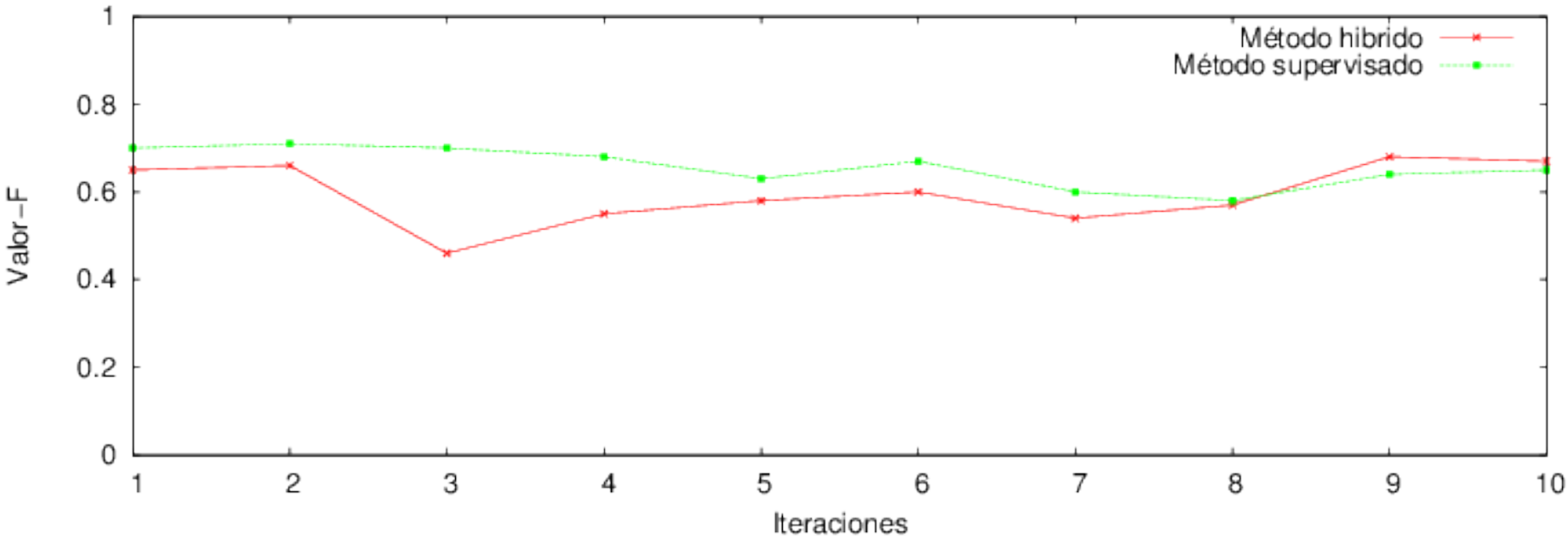


Figura 5: Valores F obtenidos en la detección de duplicados

Resultados de la detección de duplicados

Los resultados obtenidos muestran que el método supervisado tiene un mejor desempeño que el método híbrido.

Sin embargo, hay que observar el contenido los grupos resultantes del agrupamiento del método híbrido, porque estos pueden mostrar nuevas relaciones entre los reportes.

Para el tipo de documentos con los que se trabaja, los resultados son obtenidos parecen ser favorables.

Trabajos relacionados

- Agarwal, Vaithiyanathan, Sharma, and GautamShroff 2012. Detectan noticias locales reportadas en Twitter, correspondientes a "huelgas de trabajo" y "incendios en fábricas."
- Broder 2000. Detecta documentos duplicados agrupando representaciones creadas a partir de n-gramas.
- Sankaranarayanan, Samet, Teitler, Lieberman, and Sperling 2009. Implementan un sistema de generación de noticias a partir de mensajes en Twitter.
- Tao, Abel, Hauff, Houben, and Gadiraju 2013. Presentan un método de detección de duplicados para mensajes de Twitter.

Estado actual de la tesis

- Por ahora se está corrigiendo los capítulos de la tesis que ya fueron escritos.
- Todos los experimentos ya fueron realizados.
- Falta por escribir los capítulos de: introducción, resumen y conclusión.