UNIVERSIDAD CARLOS III DE MADRID

CLASIFICACIÓN AUTOMÁTICA DE VÍDEOS

Ingeniería Técnica en Informática de Gestión

Departamento de Ingeniería Telemática

Autor: David Aparicio Escribano

Tutor: Julio Villena Román

Diciembre de 2009

Título: Clasificación automática de vídeos

Autor: David Aparicio Escribano

Tutor: Julio Villena Román

EL TRIBUNAL

Presidente: Jaime José García Reinoso

Secretario: Raquel M. Crespo García

Vocal: Aitor Mendaza-Ormaza

Realizado el acto de defensa del Proyecto Fin de Carrera el día 14 de Diciembre de 2009 en Leganés, en la Escuela Politécnica Superior de la Universidad Carlos III de Madrid, acuerda otorgarle la CALIFICACIÓN de:

Fdo: Presidente Fdo: Secretario Fdo: Vocal

A mi familia, en general,
especialmente
Sofía, mi pequeña,
Sergio, mi hermano,
Florentino y María, mis padres,
Angelina y Restituto, mis abuelos.

A Julio, mi tutor.

Y yo pregunto a los economistas políticos, a los moralistas,

si han calculado el número de individuos que es necesario

condenar a la miseria, al trabajo desproporcionado, a la

desmoralización, a la infancia, a la ignorancia crapulosa, a la

desgracia invencible, a la penuria absoluta, para producir un rico.

[João Baptista da Silva Leitão de Almeida Garrett]

Este es mi pequeño trocito de proyecto que no habla de algoritmos ni números, de tablas ni gráficas, de vídeos ni textos, de pruebas y conclusiones. Y es por ello que voy a intentar darle la máxima importancia porque creo que se me lo merece.

Por el día a día en el trabajo, por tus clases y enseñanzas, por esos días esporádicos pero intensivos en el bar, por tus insistencias, por tu constancia, por tu apoyo, por las tonterías que decimos, por contar caramelos, por cerrar sitios "hablando" conmigo y compartir la siguiente mañana dura de trabajo, por tus correcciones, por tus manías, por aguantar las mías, por ver dos espacios donde los hay, y por muchas cosas más… pero ante todo, quiero darte las gracias por tu confianza y tu empeño en hacer las cosas bien, y porque sin ti este proyecto no sería, sin duda, como es. Espero haberlo hecho como querías. Muchas gracias Julio.

Que cada cual se dé por aludido y que mis amigos descubran cuando estoy hablando de ellos porque creo que todos sabréis cuando estoy hablando de vosotros. Muchas gracias a ti, que llevas conmigo desde pequeño, que has compartido tantas y tantas cosas conmigo, estudios, amigos, conocidos, bares, discusiones, charlas, sonrisas, conciertos. Por tus cabezonerías, por tus cambios de filtros, por tus "mahous y mixtas", por tus vicios en mi casa, por tus ¿has terminado el proyecto?, por tus consejos, por tus "esto que eee", por los festivales compartidos, por las pipas comidas en el césped de la universidad, por los miércoles de bares, por los martes de bares, por esos comentarios del WOW, por esos comentarios contra el WOW. Gracias por preguntarme ¿qué tal?, por aguantarme (y más en un coche, con las ventanas cerradas y en Escocia), por pedirme esa caña Vs limón hablando de juegos, por esas cartas en el parque. Por supuesto… gracias por invitarme a tu casa, en España y en Francia, en Holanda y en Suiza (me falta!!), en EEUU (todavía no!!) y en el Bierzo (que frío…).

Los amigos van y vienen y solo unos pocos duran en el tiempo, a aquellos que un día fueron mis amigos, por los momentos compartidos, gracias, a aquellos que son mis amigos y que me consideran así gracias.

A mis compañeros (y muchos de ellos amigos) de universidad, que les ha tocado avanzar, adelantarme, quedarse atrás, estudiar, no estudiar, ir a clase y demás hazañas conmigo, muchas gracias, estos años gracias a vosotros han sido mucho más bonitos, rápidos y sobre todo divertidos. A los informáticos (maldita IA), los telecos (viejunos) y los industriales (unas cuerdas + dos árboles = entretenimiento, mejor con una cerveza), y a los que habéis pasado por unas cuantas), en fin, que todos sabemos de quienes hablamos… muchas gracias chicos por todo lo compartido y por saber que a muchos de vosotros os sigo viendo pocas o muchas veces pero siempre con la misma ilusión.

Porque mi familia me ha hecho sentirme tantas veces feliz, tantas veces tranquilo. Por todas esas veces que me he sentido arropado, apoyado, mimado y atendido, gracias.

A mis tíos por vuestro apoyo y por criarme y educarme con mis padres, por dejarme crecer junto a mis primos, gracias.

A todos mis primos, desde los mayores a los pequeños (a ver cuando el siguiente :)), Nieves, Montse, Dionisilla, Enrique, Isabel (te echamos de menos, y mucho, espero que tu vida por las tierras bretañas sean cada vez mejor), Diego y Alba (a mi me toca ahora verte crecer), gracias por todos los momentos compartidos, por cuidarme y mimarme siempre, gracias.

Por cuidar de mi hermano, por hacerle feliz y por tratarme como un hermano, gracias Claudia.

Por hacerme feliz en cualquier momento, pero sobre todo en esos momentos duros que me (nos) ha tocado vivir por antojos de otros y manejos y enredos de este mundo y esta vida, gracias y mil veces gracias, abuelos, tíos, primos, familia, amigos y sobre todo Sofía, Sergio, papá y mamá, os quiero.

Porque desde que decidimos estar juntos me haces olvidarme de todo y ser feliz, pensar en ti y en mi familia y olvidarme de aquello y aquellos que no merecen la pena, por todo y por dejarme hacerte feliz, gracias Sofía.

Porque sin ti mi vida no sería la que es, porque he crecido siguiendo tus pasos, porque he crecido sabiendo que siempre has cuidado de mí, porque siempre me has apoyado, porque siempre he notado que tu mano y brazo estaban para apoyarme y para empujarme (para bien), porque eres de las mejores personas que existen, y porque eres sin duda el mejor hermano que he podido tener, gracias Sergio.

Por tu constante insistencia, por tu constante atención, porque el saber que una persona tan luchadora por los suyos es mi padre me hace sentirme el hijo más afortunado del mundo, gracias papá.

Y qué decir de ti mamá, no existe nada que hayas dicho, forma de actuar, manera de luchar, tu ímpetu en hacerme solidario, luchador, tu forma de educarme, de hacerme comprender este mundo, porque no existe nada que no pueda agradecerte, porque el saber que he tenido la suerte de tenerte como mi madre me hace sentirme feliz, gracias mamá.

Por vuestra educación, por inculcarme los valores que me habéis dado, por haberme hecho como soy, por vuestros mimos, luchas, sonrisas, broncas, besos y abrazos, nunca me cansaré de deciros gracias, gracias papa y gracias mamá.

Por último, y no menos importante, por vuestros cuidados, por los días compartidos con vosotros, por vuestras sonrisas, por vuestros besos, por vuestras broncas y riñas, por vuestros consejos y abrazos pero sobre todo por oíros decirnos siempre “sois lo único que tengo y quiero en este mundo” muchas, muchísimas gracias abuelos.

A todos vosotros, muchas gracias.

La actual tendencia a digitalizar los diferentes contenidos audiovisuales para su almacenamiento y posible explotación en medios informáticos y de telecomunicaciones está haciendo que distintas líneas de investigación se centren en procesar y analizar dichos documentos, así como buscar posibles soluciones a ciertos problemas y necesidades que traen consigo estos contenidos. La búsqueda de documentos en texto es una de las necesidades actuales mejor satisfechas mediante buscadores como Google o Yahoo en Internet, mas no es el caso de los contenidos audiovisuales. Poder consultar tanto por temática como por contenido en vídeos, audios y documentos de este estilo, abre un abanico de posibilidades bastante extenso.

La clasificación automática de contenidos audiovisuales puede ayudar a digitalizar de forma más rápida los cientos de miles de contenidos de este tipo de años atrás, consiguiendo así un ahorro de recursos y de tiempo. Puede permitir detectar vídeos con contenidos violentos, pornográficos u otros que deban ser tratados de distinta manera por ciertos usuarios.

El presente estudio pretende analizar las actuales técnicas de clasificación automática de vídeos, que distingue dos fases bien definidas, el reconocimiento automático del habla y la clasificación automática de texto. El reconocimiento automático del habla permite realizar la transcripción a texto del contenido audiovisual para posteriormente ser clasificado como un documento de texto. Las actuales líneas de investigación en clasificación automática de textos están bastante avanzadas y es por ello que el proyecto pretende seguir esta línea, convirtiendo los documentos audiovisuales en documentos de texto para, posteriormente ser procesados con técnicas de procesamiento del lenguaje natural y métodos de clasificación automática.

En definitiva, la clasificación y búsqueda de documentos audiovisuales es algo necesario en la actualidad, y aunque de momento no sea una tarea prioritaria, poco a poco debe ganar posiciones, ya que, la sociedad y en concreto el mundo que rodea Internet, requiere de documentos como vídeos y audios donde los usuarios puedan realizar consultas sobre dichos contenidos.

El proyecto que se presenta a continuación ha realizado un estudio avanzado sobre la clasificación automática de vídeos obteniendo unos resultados aceptables en un caso práctico realizado, con una precisión superior al 40% y una cobertura similar. Permite hacerse una idea de la viabilidad de estos sistemas y ofrece un estudio detallado de las actuales técnicas y líneas de investigación.

ÍNDICE DEL DOCUMENTO

ÍNDICE DE ILUSTRACIONES

ÍNDICE DE ECUACIONES

ÍNDICE DE TABLAS

ÍNDICE DE EJEMPLOS

ÍNDICE DE GRÁFICAS

ÍNDICE DE ANEXOS

CAPÍTULO 1 - INTRODUCCIÓN

1.1 INTRODUCCIÓN

La sociedad actual es altamente dependiente de los sistemas de información y del acceso a la información. La creciente predisposición a la transformación de todos los recursos referentes a información y datos a formatos digitales está tomando un papel más que relevante. Esta transformación de datos tipo expedientes, prensa, historiales, archivos legales, etc. incluso materiales multimedia tipo música, vídeos, conferencias, u otros, conlleva, mismos una vez realizada la transformación al correspondiente formato digital, la catalogación y almacenamiento ordenado de los.

El aumento de los usuarios que tienen acceso a Internet, a medios informáticos y dispositivos electrónicos donde la información es almacenada digitalmente ha implicado la investigación en técnicas y métodos que hagan de forma automática todo lo que se refiere a la transformación de los actuales documentos y materiales al formato digital, de manera que la transformación de cualquier tipo de formato, ya sea audio, vídeo o texto escrito manualmente, a un formato digital sea automática.

Como se ha comentado, de nada sirve tener todos los documentos existentes en formato digital si no se ordenan y catalogan de manera que cualquier usuario pueda tener acceso a ellos rápida y eficazmente. Imagínese una cadena de televisión que debido a la actual sociedad de la información necesita migrar todos sus programas televisivos a formatos digitales y necesita catalogar dichos programas para poder ofrecer a sus usuarios sus vídeos de manera ordenada y fácilmente accesible.

Hoy día, los materiales audiovisuales son tan importantes como los documentos textuales. Documentales, programas, series, películas, son diversos ejemplos de documentos audiovisuales que cada vez toman más y más importancia en la sociedad. La búsqueda de este tipo de documentos tanto por su contenido como por su clasificación es una necesidad que aumenta a medida que aumenta la cantidad de vídeos digitalizados y accesibles para el usuario. Sugerir vídeos similares, catalogar documentos de manera automática, realizar consultas sobre dichos vídeos son ejemplos de algunas necesidades que hoy día generan la digitalización de este tipo de documentos.

Este proyecto pretende hacer una iniciación en el mundo de la clasificación automática de vídeos y documentos audiovisuales pretendiendo explicar técnicas, algoritmos y desarrollos estudiados y servir como base de un estudio más avanzado.

1.2 MOTIVACIÓN

El presente proyecto viene motivado por una experiencia en VideoCLEF'2008 [VideoCLEF, 2008] donde el objetivo de la tarea era la clasificación automática de vídeos dados los XML con la transcripción de cada uno de ellos. La tarea fue abordada en el año 2008 con resultados importantes. Julio Villena Román y Sara Lana Serrano [Villena, Lana, 2008] investigaron y propusieron una solución basada en un motor de recuperación de información llamado Lucene [Lucene, 2009] que será la base de este proyecto.

El proyecto de fin de carrera Sistema de Indexación y búsqueda de documentos audiovisuales [Collada Pérez, 2009] ha motivado la modificación de una parte del sistema anteriormente comentado, para que, basándose en técnicas de reconocimiento automático del habla, la clasificación sea totalmente automática, es decir, dado un vídeo, realizar su transcripción audio-texto automáticamente y después clasificarlo.

De esta manera, se quiere analizar las actuales líneas de investigación en el campo de la clasificación automática de vídeos, planteando una solución y analizando los resultados obtenidos.

1.3 OBJETIVOS

El objetivo de este proyecto, es el estudio de la viabilidad, analizando sus características e implementando un sistema para poder analizar resultados, de los sistemas de clasificación automática de vídeos, dadas las características tecnológicas del momento.

Los documentos serán clasificados, en una de las categorías definidas en el sistema, aplicando diversas técnicas. El hecho de conseguir un sistema automático de clasificación evitará la intervención humana y aumentará la rapidez con que se pueden procesar este tipo de documentos.

Con la realización de este proyecto, se comprobarán y analizarán también las dificultades encontradas en la implementación de un sistema de clasificación automática donde la naturaleza de los documentos es de tipo vídeo (concretamente tomando el audio) y estos han sido elaborados por usuarios cualesquiera de Internet, los cuales pueden decidir libremente cualquier clasificación para cada vídeo.

Para resumir, se podría decir que los principales objetivos buscados por el proyecto son los siguientes:

- Construir un corpus de vídeos adecuado para la evaluación de un sistema de clasificación automática de vídeos, extrayendo la información relevante de los mismos, mediante el reconocimiento automático del habla y el procesamiento del

lenguaje natural. En concreto, en este proyecto, se obtendrán vídeos de YouTube [YouTube, 2009] de diferentes categorías.

- Adquirir un conocimiento suficiente para poder dar una posible solución a la clasificación automática de vídeos estudiando las actuales técnicas de clasificación automática de vídeos.
- Realizar un conjunto de pruebas suficientemente robusto como para poder obtener unas conclusiones de viabilidad, dificultad y eficiencia de este tipo de sistemas.

1.4 DESCRIPCIÓN DEL DOCUMENTO

El presente documento se estructura de la siguiente manera:

- **Capítulo 1 - Introducción:**

  Presente capítulo, donde se presentan los fundamentos y motivaciones del proyecto, centrándose concretamente en las motivaciones y fundamentos que tienen los sistemas de clasificación automática de vídeos y la necesidad de los mismos en la actual sociedad. Se enumeran también los objetivos del proyecto y se describe la estructura de la memoria.
- **Capítulo 2 - Estado del arte:**

  Visión general de los sistemas de clasificación automática de vídeos, las actuales técnicas, líneas de investigación y campos relacionados con este mundo. Se detallan fases del diseño de estos sistemas, se analizan algoritmos de clasificación, y se explican y detallan actuales sistemas de reconocimiento automático del habla y de clasificación automática de documentos.

  Por otra parte, también se explicará la arquitectura típica de un clasificador, así como algunas soluciones ya realizadas en otros trabajos.
- **Capítulo 3 - Diseño e implementación de un sistema de clasificación de vídeos:**

  Se explicará las decisiones tomadas y los procesos realizados del clasificador que se ha implementado y que ha servido como base para las pruebas realizadas. Se darán razones y explicaciones por las que se han elegido ciertos criterios y no otros.
- **Capítulo 4 - Evaluación:**

  Explicación de los resultados obtenidos, explicando las distintas pruebas y criterios. Paralelamente se comentan conclusiones de los mejores resultados. Por último, un breve resumen y comparativa de los mejores resultados obtenidos.
- **Capítulo 5 - Conclusiones y trabajos futuros:**

  En este último capítulo, se presentan las conclusiones alcanzadas tras la realización del proyecto y el análisis de los resultados obtenidos. Por otra parte, se presentan posibles líneas de investigación en las que seguir trabajando con el fin de obtener mejoras en el sistema y ahondar más en el ámbito de la clasificación automática de vídeos.
- **Anexos**
- **Referencias**

CAPÍTULO 2 - ESTADO DEL ARTE

2.1 INTRODUCCIÓN

La clasificación automática de vídeos es un campo que requiere de diversas técnicas y líneas de investigación. Es un tema incipiente, por lo que actualmente existen pocos estudios sobre este mundo.

Este capítulo pretende introducir al lector en las áreas más afines y que más relación tienen con la clasificación automática de vídeos. Técnicas de procesamiento de documentos, reconocimiento automático del habla, algoritmos de clasificación de textos, y otros serán los apartados de este capítulo.

La ingeniería lingüística pretende representar la información de manera que las máquinas puedan entenderla. Se plantearán por lo tanto problemas como la ambigüedad y otros que atañen al proyecto.

La clasificación automática de documentos y el reconocimiento automático del habla son los puntos más importantes del capítulo.

El capítulo se terminará hablando de otros trabajos parecidos y explicando la arquitectura típica de un clasificador automático de vídeos.

2.2 LA INGENIERÍA LINGÜÍSTICA

2.2.1 INTRODUCCIÓN

La Ingeniería Lingüística proporciona investigación y desarrollo, en general, medios para ampliar y mejorar la utilización del lenguaje natural, para conseguir potenciar su utilización en los sistemas informáticos, asimilando, analizando, seleccionando y presentando la información de manera que las máquinas puedan llegar a "entender" e interpretar el lenguaje que emplean los seres humanos [Llisterri, 2008].

2.2.2 CAMPOS DE ESTUDIO

Existen diversos campos de estudio en la ingeniería lingüística. La cotidianidad del uso del lenguaje y el creciente interés que está despertando el procesamiento del mismo, sumado

a la inminente investigación y desarrollo de aplicaciones, han dado como resultado ciertos campos bien definidos. Se podría decir que los campos de estudio más importantes son:

- Revisión lingüística de textos.
- Recuperación de información.
- Extracción de información, resúmenes y clasificación.
- Reconocimiento y síntesis de voz.
- Traducción automática.
- Generación automática de texto.

2.2.3 NIVELES DEL PROCESAMIENTO DEL LENGUAJE NATURAL

Típicamente se consideran seis niveles en el proceso de análisis del lenguaje natural y su transformación a un lenguaje "entendido" por la máquina.

La síntesis del lenguaje consiste en unir todos los niveles para adquirir el significado del mismo. Por otra parte, el análisis del lenguaje consiste en la distinción y separación de cada nivel del lenguaje para realizar un estudio minucioso. Los niveles del lenguaje son [Villena Román, 2008]:

- **Nivel Fonológico:** Se encarga de la conversión de la voz a texto. Como es evidente es quizás el lugar donde la ambigüedad da más juego. Entonaciones, intenciones, letras mudas, entre otros, son los grandes problemas de este nivel.

Hola / ola

Ejemplo 1 - Nivel Fonológico

- **Nivel Morfológico:** El estudio morfológico se centra en la estructura de las palabras, etiquetando así las palabras para generar una estructura etiquetada por lemas y categorías gramaticales. En el siguiente ejemplo, el lema de niño es "niñ" al igual que el de niña.

Niño y niña-> "o" y "a" masculino o singular ; "niñ" lema

Ejemplo 2 - Nivel morfológico

- **Nivel sintáctico:** El nivel sintáctico hace referencia a la clasificación de las palabras según su orden para formar oraciones y expresar contextos. En este sentido, la ingeniería lingüística genera estructuras representando agrupaciones de palabras y relaciones. ¿Quién es el cazador? o ¿quién es el cazado? es un ejemplo de nivel sintáctico, es decir, la estructura de la frase.

El lobo trató de cazar a Caperucita

Ejemplo 3 - Nivel Sintáctico

- **Nivel Semántico:** Es referente al significado semántico de la frase, es decir, al conjunto de unidades léxicas de una lengua que comprende términos ligados entre sí por

referirse a un mismo orden de realidades o ideas. Existen reglas y otros recursos que consigue acercarse e intentan simular este nivel mas la ambigüedad del lenguaje hace de esto una más que ardua tarea, por ejemplo:

Pasé delante del banco - ¿qué banco?

Ejemplo 4 - Nivel Semántico

- **Nivel Pragmático:** Cuando la intención de una frase es totalmente distinta del significado real de la misma.

¿Puedes pasarme la sal?

- *Significado literal de frase = Sí o no*
- *Significado real de frase = El individuo que pregunta quiere que le den la sal.*

Ejemplo 5 - Nivel Pragmático

- **Nivel de Integración del discurso:** Este nivel hace referencia al significado del lenguaje en un contexto concreto, por ejemplo

Me dijo que se lo daría

- *Significado de frase aislada = La frase "no tiene" significado*
- *Significado en contexto = En un contexto, los hablantes saben qué es lo que le iba a dar*

Ejemplo 6 - Nivel de Integración

2.2.4 LA AMBIGÜEDAD Y LOS PROBLEMAS DEL PROCESAMIENTO DEL LENGUAJE NATURAL

El procesamiento del lenguaje natural no es algo trivial ni mucho menos. Los problemas y dificultades a resolver son considerables.

Se podría decir que existen una serie de dificultades que, quizás por su relevancia, llaman más la atención, son citados a continuación.

- **La ambigüedad:** Según la RAE [RAE, 2009] dicho especialmente del lenguaje: Que puede entenderse de varios modos o admitir distintas interpretaciones y dar, por consiguiente, motivo a dudas, incertidumbre o confusión.

  Quizás la razón por la que la ambigüedad es tan complicada de tratar es porque en la mayoría de los casos, requiere un análisis previo de los niveles superiores. Por ejemplo, si se está analizando en el nivel fonológico el sonido */ola/* es necesario hacer un estudio morfológico, sintáctico y quizás semántico ya que en ningún caso será posible saber su verdadera forma hasta no haber analizado los niveles superiores.

- **La Semántica:** Es uno de los grandes problemas, está ligada totalmente a la ambigüedad y depende del dominio concreto.

  Existen recursos semánticos que son capaces de crear redes semánticas entre conceptos, como por ejemplo:

  - Sinonimia: Palabra que tiene el mismo significado o parecido que otro, es decir, una relación de semejanza entre dos palabras. Por ejemplo, *comprobar y verificar*.
  - Hiponimia: Cuando el significado de una palabra está incluida en otra, por ejemplo, gorrión *con respecto a pájaro*.
- **Complejidad de niveles superiores:** Los niveles por encima del nivel semántico son extremadamente complejos debido a la ambigüedad interna de los mismos, es decir, conceptos como la intención (nivel pragmático) o el contexto (nivel de integración del discurso) no pueden ser analizados ni asimilados por una máquina.
- **Diferencias entre lenguas:** Para comprender el problema que acarrea la diferencia entre lenguas, es importante definir lengua y lenguaje [RAE, 2009]:
  - Lengua:
    - Sistema de comunicación verbal y casi siempre escrito, propio de una comunidad humana.
    - Sistema lingüístico cuyos hablantes reconocen modelos de buena expresión.
    - Sistema lingüístico considerado en su estructura.
  - Lenguaje:
    - Conjunto de sonidos articulados con que el hombre manifiesta lo que piensa o siente.
    - Manera de expresarse. *Lenguaje culto, grosero, sencillo, técnico, forense, vulgar.*
    - Estilo y modo de hablar y escribir de cada persona en particular.

  La gran variedad de lenguas, de formas de hablar, la diferencia entre las gramáticas, todo esto hace que el procesamiento del lenguaje deba fijar un idioma concreto y ambientar y dirigir cada aplicación, cada desarrollo en el ámbito de esta única lengua.

2.3 RECUPERACIÓN DE INFORMACIÓN

La recuperación de información (RI*, Information Retrieval*, IR en inglés) [López Herrera, 2005] se define como el problema de la selección de documentos en respuesta a consultas o demandas de información por parte de un usuario. Los sistemas de RI utilizan bases de datos compuestas por documentos y procesan las consultas de los usuarios permitiéndoles acceder a la información relevante en un intervalo de tiempo apropiado.

La indexación es la solución a las necesidades de los sistemas de RI. El objetivo de estos sistemas es encontrar, de manera sencilla, automática y mediante el uso de consultas en

lenguaje natural, el mayor número posible de objetos relevantes. La indexación es la representación apropiada de los documentos seleccionando aquellos términos que mejor caracterizan a dichos documentos, facilitando así la RI.

El lenguaje mantiene una relación estrecha con la construcción del conocimiento y desempeña un papel crítico en las operaciones de RI. Los sistemas de RI basados en texto son aplicaciones cuyo objetivo es resolver el problema de la búsqueda de texto en bases de datos.

De esta manera, se muestra en la Ilustración 1 la RI desde un mecanismo de almacenamiento en respuesta a consultas realizadas por un usuario [Luque Rodríguez, 2006].

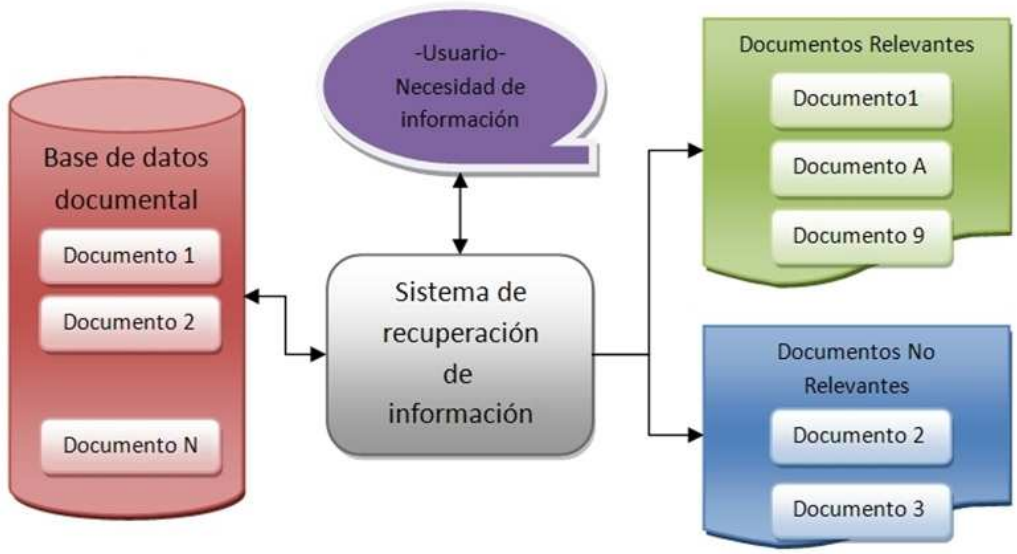


Ilustración 1 - Recuperación de Información

Un sistema de estas características debe soportar además una serie de operaciones básicas sobre los documentos almacenados en el mismo, como son: introducción de nuevos documentos, modificación de los que ya estén almacenados y eliminación de los mismos.

Centrándose en la recuperación de información textual (los vídeos serán transcritos), que es la que atañe a este proyecto, los sistemas de RI utilizan términos índice (palabras clave) para expresar los contenidos de los documentos. Estos términos poseen significado propio por lo que representan un concepto.

El proceso de extracción de claves conlleva los siguientes pasos [Paz, 2007]:

- **Conversión de formato**: Convertir el documento en texto que pueda analizar y tratar el sistema, pero manteniendo la información útil que exista (negritas o cursivas, por ejemplo).
- **Análisis léxico del texto**: Transformar el texto en un conjunto de *tokens* (palabras o expresiones multipalabra). Tratamiento concreto de números, fechas, mayúsculas, nombres propios, etc.

- **Eliminación de palabras vacías (*stopwords*)**: Palabras muy frecuentes o muy poco frecuentes.
- **Lematización o extracción de raíces:** es decir convertir palabras como *niño* y *niña* en *niñ* para poder recoger juntas todas las que pertenezcan a la misma familia.
- **Selección de términos:** Los que serán considerados términos índice. Suele ser útil realizar un análisis morfosintáctico para seleccionar ciertas categorías de palabras, como nombres, adjetivos y verbos, puesto que son las que aportan un mayor contenido semántico.
- **Expansión de claves**: Inclusión en el índice de términos relacionados con el contenido semántico del documento, por ejemplo, sinónimos o términos traducidos a otros idiomas.

El propósito de la RI, como queda patente, es crear mecanismos para localizar información en grandes colecciones de documentos en formato electrónico. Para obtener esta información, normalmente, se comienza con la selección de términos o expresiones que se consideran importantes para la búsqueda. Después se generan índices que reflejan la frecuencia de aparición de dichos términos. Una vez realizada la consulta, el sistema devuelve un subconjunto de documentos relevantes.

En la actualidad, la RI ha cobrado mayor importancia debido a los muchos tipos de información disponibles; concretamente, la cantidad de información del sector audiovisual se ha masificado en los últimos años, adquiriendo mayor importancia la necesidad de indexarla.

El avance de servicios multimedia ocurrido con el progreso tecnológico y la posibilidad de compartir y distribuir datos a través de las redes de comunicación han acentuado la importancia de herramientas para la recuperación de información multimedia. Las bases de datos de imágenes se emplean en un vasto abanico de áreas como son el entretenimiento, el arte, la publicidad, la medicina y la industria entre otros. En todos estos contextos, el problema principal está relacionado con la necesidad de un acceso eficiente a la información [Collada Pérez, 2009].

2.4 SISTEMAS DE CLASIFICACIÓN AUTOMÁTICA DE VÍDEO

2.4.1 INTRODUCCIÓN

La clasificación automática de vídeos requiere plantear y estudiar varios campos de investigación totalmente distintos.

Actualmente es un problema poco planteado y poco estudiado, mas las actuales Tecnologías de la Información y las Comunicaciones (TIC) se enfrentan al filtrado, selección y gestión de la información. Los contenidos multimedia tipo vídeos o audios, tarde o temprano deberán pasar por los mismos procesos (selección, filtrado y gestión) requiriendo entonces de ciertas técnicas de recuperación que permitan hacer estos procesos de manera automática.

Actualmente hay una tarea específica de CLEF [CLEF, 2009] para estos sistemas [VideoCLEF, 2008].

CLEF, acrónimo de *Cross Language Evaluation Forum,* es el foro europeo más importante para la evaluación de sistemas de recuperación multilingüe y multimedia. CLEF está financiado desde el 2000 por la Unión Europea y está coordinado por el *Istituto di Scienza e Tecnologie dell'Informazione*, *del Consiglio Nazionale delle Ricerche (CNR)* en Italia. Es un foro competitivo en el que diferentes grupos envían sus resultados para una serie de tareas y se comparan las diferentes técnicas entre sí, en términos de precisión y *recall* principalmente.

El aumento del número de fuentes y documentos audiovisuales sumado a la actual tendencia de digitalizar todos los contenidos, hace más que patente la necesidad de estudiar métodos y técnicas de clasificación automática de estos contenidos.

La clasificación automática de vídeos requiere de técnicas de reconocimiento del habla y técnicas de clasificación automática de documentos [Perea-Ortega, 2008] aunque existen otros sistemas basados en fotogramas [Lu, 2002] [Jeong, 2002].

José Manuel -Ortega junto con otros compañeros [Perea-Ortega, 2008] han creado un sistema de clasificación automática de vídeos basándose en las transcripciones de los mismos. En su proyecto SINAI [Perea-Ortega, 2008], con el que han participado dos años en VideoCLEF, han conseguido buenos resultados. SINAI se basa en la recolección de documentos de las diferentes categorías basándose en el motor de búsqueda Google [Google, 2009], para posteriormente indexar dichos documentos.

Estos documentos indexados serán la base de entrenamiento para el sistema. Para posteriormente crear la consulta al sistema, SINAI utiliza la lematización y la eliminación de las palabras vacías y así poder clasificar un vídeo. Estas técnicas serán estudiadas y explicadas más adelante.

El sistema de RI que utiliza SINAI es LEMUR [LEMUR, 2009], un *software OpenSource* desarrollado, en colaboración, por la Universidad Carnegie Megion y la Universidad de Massachusetts. Está desarrollado en C++ y permite indexar y buscar en documentos.

Por otra parte, existen, como se ha dicho, sistemas basados en fotogramas. Un buen ejemplo de estos sistemas es el de Chen Lu [Lu, 2002]. Extrae las características de los fotogramas de los vídeos, para crear fotogramas clave y extraer de ellos ciertas particularidades de las imágenes. Posteriormente se crea una agrupación jerárquica de los vídeos y se procesa. Este sistema se basa en los modelos ocultos de Markov para encontrar la categoría de los vídeos.

El sistema presentado por Jeong [Jeong, 2002] pretende clasificar un vídeo como censurable o no. Utiliza 3 analizadores digitales para obtener características de los fotogramas y obtener así descriptores visuales. Se basa en las máquinas de vectores soporte y da unos resultados bastante aceptables.

Incluso existen sistemas que mezclan ambas técnicas. Ide [Ide, 1999] propone un sistema que analiza primeramente las características de los fotogramas y después un análisis semántico de los términos que representan al vídeo.

El presente capítulo pretende hacer una explicación de la arquitectura típica de un sistema de clasificación automática de vídeos explicando y analizando técnicas y procedimientos tanto de reconocimiento del habla como de clasificación automática de textos.

Para poder comprender bien el funcionamiento típico de un clasificador automático de vídeos es necesario separar claramente dos partes, por un lado la clasificación automática de documentos, que será el grueso principal del programa y por otro el reconocimiento automático del habla que será la parte que condicionará en gran medida la calidad de los documentos obtenidos.

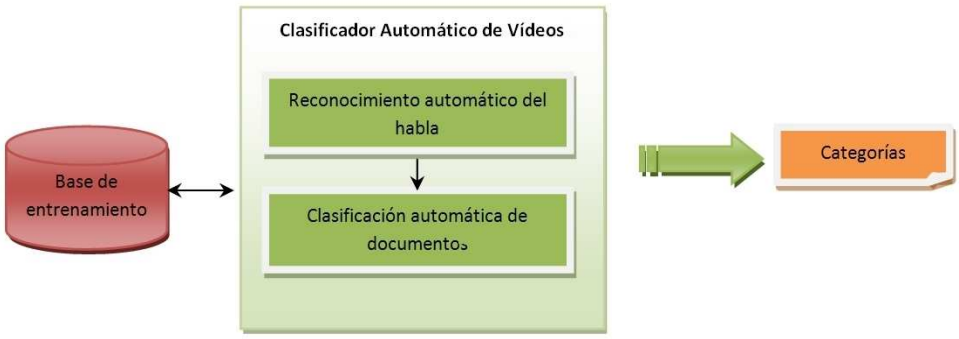


Ilustración 2 - Clasificador automático de vídeos

El reconocimiento automático del habla consiste en hacer la transcripción audio-texto de un audio, es decir, dado un audio identificar las palabras del mismo, en el orden y tiempo correctos (será explicado más detalladamente en los siguientes apartados).

Hoy día, la clasificación automática de vídeos se realiza mediante el audio ya que las técnicas conocidas de similitud de imágenes no están lo suficientemente avanzadas como para poder realizar una clasificación basada en fotogramas. Por tanto, y como se viene diciendo, se realiza un proceso de extracción del audio ya que será el audio del vídeo el que representa a dicho vídeo [Ide, 1999].

Por otra parte, la clasificación automática de documentos consiste en asignar categorías, pertenecientes a un conjunto previamente dado, a documentos escritos en lenguaje natural. A continuación se describirá cada uno de los procesos por separado.

2.4.2 RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Como se ha comentado anteriormente, se denomina reconocimiento automático del habla al proceso automático por el cuál un sistema es capaz de identificar una señal de voz producida por un individuo o conjunto de ellos. Esta señal es sometida a procesos de

digitalización con el fin de obtener características y propiedades que permitan estudiar y analizar su comportamiento y desarrollar procesos para tratar dicha señal y reconocerla.

La estructura típica de un sistema de reconocimiento automático del habla puede apreciarse en la siguiente figura.

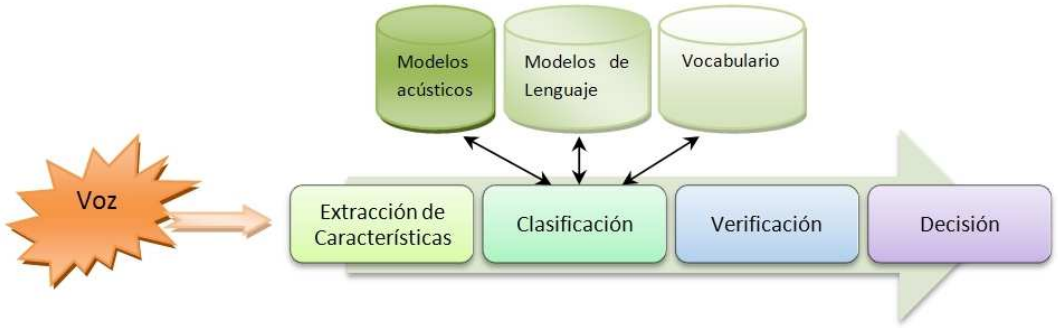


Ilustración 3 - Reconocimiento automático del habla

El reconocimiento de la señal de la voz es el primer paso, como bien se muestra en el esquema, y con ello lo que se pretende es obtener valores que contengan información sobre la señal emitida por el individuo, esta información es aquella que permite al sistema reconocer el mensaje. El proceso de reconocimiento automático del habla es computacionalmente complejo, además de bastante costoso, por ello es recomendable que la información obtenida en el análisis de la señal sea la meramente imprescindible.

La voz se limita en banda y se digitaliza, para posteriormente dividir dicha señal en segmentos de duración fija y solapados entre sí. Cada uno de estos segmentos comentados anteriormente son tratados en un análisis frecuencial de la señal para dar un conjunto representativo de parámetros.

El análisis frecuencial puede realizarse mediante diferentes métodos [Moreno, 2009]:

- Filtrar la señal en distintas bandas frecuenciales y calcular la energía en cada banda. Cada segmento quedará representado por n valores de energía correspondientes a una banda.
- Con técnicas de predicción lineal, calcular el espectro envolvente que permita extraer para cada segmento los parámetros que representan el modelo articulatorio y que lleven información del mismo. Una vez que se ha parametrizado la señal, se realiza una estimación de las características dinámicas del espectro de la señal de voz, tanto una evolución temporal del espectro como de la energía. Una vez recogida toda la información comentada anteriormente se procede al reconocimiento de la palabra.

Para dicho reconocimiento el sistema inicia una búsqueda donde encontrará qué palabra se parece más a la reconocida, gracias a los parámetros obtenidos y los diccionarios de términos correspondientes. Es importante tener en cuenta que dependiendo del tipo de

reconocimiento (continua o no) el reconocimiento en sí puede complicarse ya que entran factores como dónde acaban y dónde empiezan las palabras. A partir de este punto es importante definir los siguientes campos de conocimiento:

- **Modelos acústicos:** Permiten establecer la distribución de los parámetros acústicos de los fonemas, es decir, centran su esfuerzo en el correcto modelado de las señales a reconocer [Nogueiras, 1999].

[f] es labiodental fricativa sorda.

Ejemplo 7 - Modelos acústicos

- **Modelos del lenguaje:** se refiere a la ordenación de las palabras en una lengua, es decir, dan información sintáctica y semántica al sistema [Olaso, 2003].

"pescado fresco" / "pescado estrafalario".

Ejemplo 8 - Modelos del Lenguaje

Con estos dos campos lo que se pretende es crear una expectativa de lo que se está diciendo, es decir, intentar intuir qué es lo que debería ser, de tal manera que con la señal de entrada y estas expectativas se pueda construir una hipótesis de aquello que un individuo está diciendo [Gavaldà i Camps, 2009].

Se realizan también procesados para la comprensión del lenguaje natural donde se pretende dar una representación semántica y sintáctica de la frase a reconocer.

2.4.2.1 CLASIFICACIÓN DE LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Para diferenciar los sistemas de reconocimiento del habla es necesario observar el uso que se va a dar a la aplicación ya que esto, en gran medida, permitirá saber la precisión que ha de dar el sistema elegido. La tabla que se muestra a continuación muestra la relación entre el tipo de lenguaje y el error común en los sistemas de reconocimiento del habla [Collada Pérez, 2009].

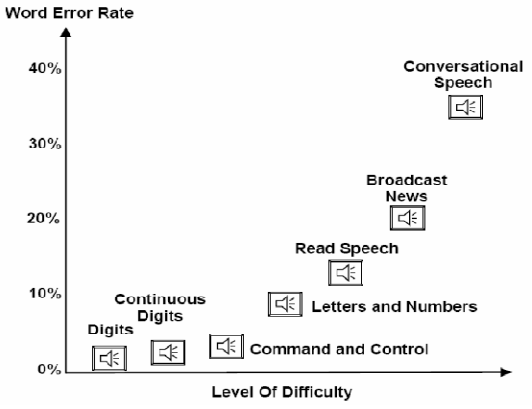


Ilustración 4 - Tasa de error y dificultad en sistemas de reconocimiento del habla

2.4.2.1.1 SISTEMAS DE RECONOCIMIENTO DEL HABLA Y EL HABLANTE

Típicamente los sistemas pueden ser dependientes, independientes o incluso adaptados al hablante. Esto quiere decir, que si un sistema es dependiente del hablante funcionará de forma óptima para dicho hablante, es decir, que será entrenado para que su rendimiento y precisión sea máximo con dicho hablante. Estos sistemas suelen ser más baratos, fáciles de desarrollar y más precisos, en cambio, pierden flexibilidad.

Existen también sistemas capaces de adaptarse a nuevos hablantes, requieren de técnicas que permitan encontrar un punto medio entre los hablantes dependientes e independientes de él.

Por otro lado, existen los sistemas independientes del hablante, además de su mayor dificultad a la hora del desarrollo, por lo general su precisión es menor a cambio de una mayor flexibilidad.

Los sistemas independientes del hablante son los que más atañen a este proyecto, ya que un sistema de reconocimiento automático del habla ambientado a vídeos de distintas categorías, donde interactúan diferentes situaciones, locutores, ambientes es el único que puede permitir a un sistema de clasificación automática de vídeos hacer su trabajo.

Son aquellos sistemas que están preparados para reconocer el lenguaje, la voz, de cualquier hablante. Un sistema de reconocimiento automático del habla que sirva para cualquier hablante de un idioma o dialecto, su base de datos de aprendizaje deberá contener las voces de un número elevado de locutores [Nadeu, 2001].

Los sistemas independientes del hablante permiten, por tanto, reconocer el habla de una conversación con distintos locutores, y son por lo tanto los necesarios para los sistemas de clasificación automática de vídeos.

2.4.2.1.2 EL VOCABULARIO EN LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

El vocabulario de un sistema comprometerá al mismo en ciertos aspectos, entre ellos la complejidad, la precisión del sistema y los requerimientos del procesado.

El tamaño del mismo suele depender el uso de la aplicación, se suele seguir la siguiente clasificación:

- vocabulario pequeño - decenas de palabras.
- vocabulario mediano - cientos de palabras.
- vocabulario grande - miles de palabras.
- vocabulario muy grande - decenas de miles de palabras.

2.4.2.1.3 EL RECONOCIMIENTO CONTINUO O DISCRETO

Diferenciar entre el tipo de reconocimiento (continuo o discreto) permite conocer los requerimientos del sistema y adaptar así la aplicación.

Un reconocimiento discreto implica que la señal a reconocer son palabras simples con pausas entre palabras y que la pronunciación de las mismas no implica dependencia entre ellas. Suele ser un reconocimiento más sencillo.

Un reconocimiento continuo pretende reconocer palabras y expresiones en frases continuas. Para ello necesita encontrar el inicio y fin de cada palabra, además la dependencia en la pronunciación suele venir por las palabras anteriores de manera que las palabras no suenan siempre igual.

2.4.2.2 LIMITACIÓN DE LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Como todo sistema, tiene sus limitaciones, entre los factores más importantes se encuentran el ruido, factores ambientales, conversaciones simultáneas, intenciones y emociones del hablante, etc.

Cuando el sistema pretende captar la señal ha de limpiar el ruido. Para ello existen técnicas que eliminan aquellas partes de la señal que se asemejen a los patrones de ruido. Es importante que en esta primera fase el sistema sea robusto ya que de esta señal dependerá todo el resto de procesos posteriores.

Por otro lado, los sistemas de reconocimiento del habla se enfrentan a la resolución de las conversaciones múltiples, algo que hasta ahora no se ha resuelto. No se han conseguido técnicas que, como el oído humano, puedan discriminar y seguir una conversación entre varias. No obstante, existen técnicas de cancelación a partir de señales tomadas simultáneamente.

2.4.2.3 APLICACIONES DEL RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Las aplicaciones para las que puede usarse un reconocedor automático del habla son distintas y variadas, entre ellas se ha intentado destacar ciertas aplicaciones o campos que se han creído más importantes y con más fuerza hoy día.

Es evidente que donde alcanza un gran potencial el reconocimiento del habla es en el uso como interfaz entre máquina e individuo. Desde esta posición son varias y distintas las ramas que crean aplicaciones y servicios con esta tecnología.

Por otra parte, y es campo que atañe a este proyecto, la transcripción a texto del lenguaje natural que, como se viene indicando en este mismo documento, permite la indexación y catalogación de documentos audiovisuales.

Por tanto, se podría decir que quizás las aplicaciones más importantes o más relevantes hoy día son las siguientes:

- Sistemas de control e interacción con máquinas. Por ejemplo contestadores automáticos.
- Domótica y accesibilidad para disminuidos físicos. Control de aparatos mediante el habla.

- Transcripción del habla. Consultas y accesos a bases de datos, catalogación automática de documentos audiovisuales, etc.

2.4.2.4 SISTEMAS ACTUALES DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Hoy día existen diversos sistemas de reconocimiento automático del habla, a continuación se van a citar unos cuantos que se han considerado importantes en otros proyectos [Collada Pérez, 2009]

- **Dragon NaturallySpeaking [Dragon, 2009]:**

  Uno de los primeros programas en realizar reconocimiento de voz en ordenadores personales. Tiene tres funcionalidades principales:

  - Dictado: el lenguaje hablado es transformado a texto.
  - Comandos de control: el lenguaje hablado es reconocido como un comando para controlar el ordenador.
  - Sintetizador de voz: transformación del texto escrito en voz.

  Actualmente el sistema aprende de forma interactiva las características de habla del usuario y reconoce palabras aisladas con separaciones entre palabras de un cuarto de segundo pudiendo llegar a crear un texto a una velocidad de hasta 120 palabras por minuto.

- **ViaVoice [ViaVoice, 2009]:**

  Este programa reconoce listas de vocabulario que excede las 200.000 palabras en tiempo real y en diferentes idiomas. ViaVoice también proporciona un conjunto de herramientas de uso intuitivo potenciado por la tecnología de Eclipse.

  Dispone de una arquitectura completamente integrada que proporciona reconocimiento automático del habla, síntesis de voz y otras tecnologías. Una arquitectura sencilla que permite la implementación de cualquier tipo de sistema mediante la extensión de las capacidades de la plataforma.

  Es un sistema de reconocimiento de habla versátil capaz de funcionar en diversos procesadores y sistemas operativos, y se encuentra en una gran variedad de idiomas. Cuenta al mismo tiempo con un amplio vocabulario, el tamaño del vocabulario que puede reconocerse ha crecido en un factor de 25 en los últimos cuatro años, llegando a ser superior a 200.000 palabras en tiempo real.

  El motor de reconocimiento está basado en pequeñas unidades de audio llamadas fonemas. El modelo basado en fonemas utiliza estados finitos para conseguir una alta precisión y un sistema robusto frente al ruido en el reconocimiento de habla continuo mejorando la detección de voz y silencio.

- **Media Mining Indexer [Media Mining Indexer, 2009]:**

  Es el sistema que se ha utilizado en el proyecto y es explicado en detalle en el Capítulo 3 -la característica más importante es su independencia del hablante y es por ello la razón de su elección.

- **Sphinx [Sphinx, 2009]:**

  Este sistema fue desarrollado por la Universidad Carnegie Mellon y con colaboraciones de Sun Microsystems y de los laboratorios de investigación de Mitsubishi Electric.

Utiliza Modelos Ocultos de Markov y funciones de densidad probabilísticas. Existen varias versiones disponibles que se pueden emplear de manera gratuita.

El funcionamiento de Sphinx carece de la precisión y eficiencia de los sistemas comerciales, es más el modelo disponible en español agrava aún más las desventajas mencionadas.

- **Julius [Julius, 2009]:**

  Es un motor de reconocimiento de habla continuo de alto reconocimiento que cuenta con dos grandes vocabularios para realizar la conversión de voz a texto. Es un sistema de reconocimiento automático del habla de código abierto.

  Julius está basado en trigramas y utiliza modelos ocultos de Markov, se puede utilizar en aplicaciones en las que sea necesario obtener la transcripción de documentos de audio en tiempo real e incluso como herramienta de dictado. Adopta modelos acústicos y diccionario de pronunciaciones en formato HTK y modelos de lenguaje basados en trigramas en formato ARPA.

  Se diseñó inicialmente para realizar reconocimiento automático del habla en japonés actualmente existe un modelo acústico en inglés, además aunque está diseñado para plataformas Unix también funciona correctamente en Windows.

2.4.3 CLASIFICACIÓN AUTOMÁTICA DE DOCUMENTOS

2.4.3.1 INTRODUCCIÓN

La clasificación automática de documentos es de carácter híbrido [Sánchez Jiménez, 2007] ya que viene dado por el uso de principios y metodologías de la Inteligencia Artificial para conseguir objetivos de la RI, es un campo de estudio multidisciplinar, donde deben involucrarse tanto la Lingüística Documental como la Documentación.

Se podría decir que el proceso de clasificación es [Chan, M.L, 1981] *"el acto de organizar el universo del conocimiento en algún orden sistemático. Ha sido considerada la actividad fundamental de la mente humana. El acto de clasificar consiste en el dicotómico proceso de distinguir cosas u objetos que poseen cierta característica de aquellos que no la tienen y agrupar en una clase cosas u objetos que tienen la propiedad o característica en común".*

Para comprender bien en qué consiste una clasificación automática es preciso comprender en qué consiste una clasificación manual. Dicha clasificación consiste en un análisis del contenido, esquematizar dichos contenidos y contrastar el tema principal o materia con las clases previamente asignadas.

Debido a la gran cantidad de documentación, de información disponible en la actualidad, es necesario el estudio y el aprendizaje de técnicas de clasificación automática que permitan crear sistemas potentes y eficaces. La gestión de la información es una tarea de carácter obligatorio hoy día y es por ello el creciente interés que ciertos sectores están teniendo en estas, relativamente nuevas, áreas de investigación.

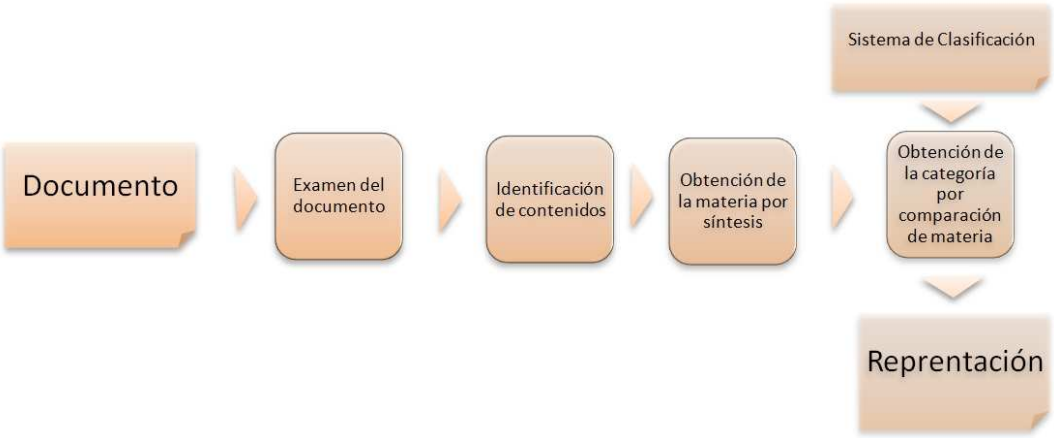


Ilustración 5 - Esquema de Clasificación Manual

No obstante, la clasificación automática depende de ciertos criterios que han de explicarse para poder comprenderla. A grandes rasgos se podría decir que la clasificación automática de textos consiste en un conjunto de algoritmos, técnicas y sistemas capaces de asignar a un documento una o varias categorías o clases según su afinidad temática. Como se ha comentado anteriormente utiliza técnicas de aprendizaje automático y técnicas de procesamiento del lenguaje natural.

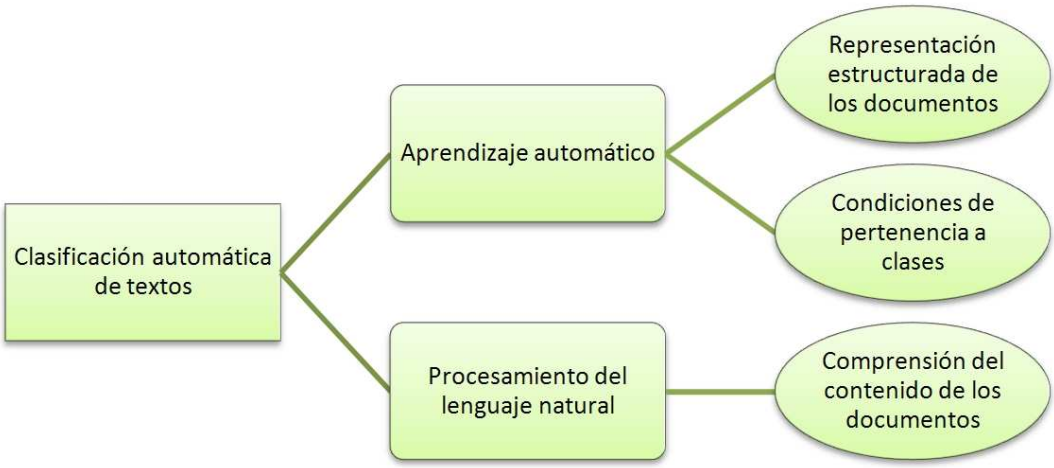


Ilustración 6 - Clasificación Automática de Textos

El comportamiento típico de un clasificador automático está dividido en dos fases, la fase de entrenamiento, donde el sistema pretende aprender cómo son las categorías donde en un futuro deberá clasificar los documentos y la fase de clasificación, donde el sistema, mediante técnicas y algoritmos matemáticos, indica que clase o categoría es la más parecida para un documento dado.

El *set* de entrenamiento o conjunto de entrenamiento de un sistema se crea escogiendo documentos que representan a cada una de las categorías, estos documentos,

deben ser clasificados por expertos en la materia. Este conjunto de documentos, servirá tanto para el entrenamiento del sistema como para su evaluación. Típicamente se hace una relación 80% a 20%, es decir, de estos documentos escogidos por expertos que representan a cada clase o categoría, se escogen un 20% para realizar las pruebas (conjunto de test) y el resto para realizar el entrenamiento del sistema.

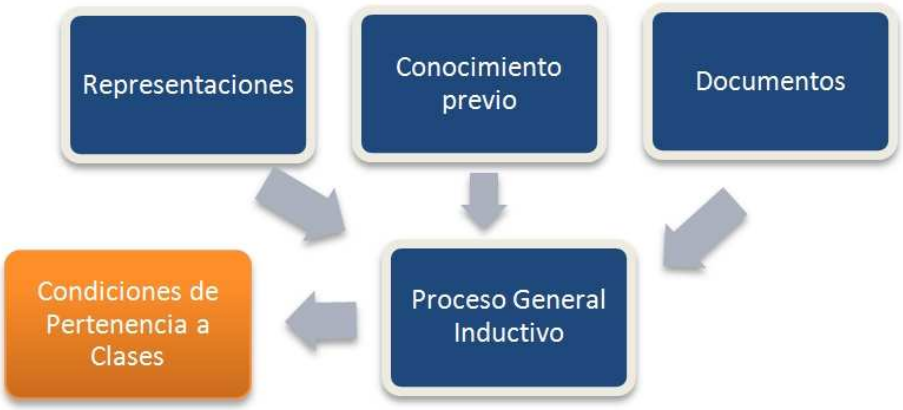


Ilustración 7 - Fase de aprendizaje en la Clasificación Automática.

El anterior y siguiente esquema (Ilustración 7 e Ilustración 8) resumen y representan de manera sencilla lo anterior.

Dentro del aprendizaje automático se requiere de una representación estructurada de los documentos que dependerá del modelo matemático que se utilizará para la clasificación.

La más frecuente es el modelo de espacio vectorial, donde cada documento se convierte en un vector de palabras asignándolas y ponderándolas dependiendo de ciertas características que resumen su importancia en el texto.

Ilustración 8 - Fase de decisión en la Clasificación Automática.

Fíjese en la relación entre una ilustración y otra, es decir, las condiciones que requiere la clasificación son dadas por el aprendizaje y entrenamiento del sistema, lo que quiere decir, que en gran medida, los resultados obtenidos estarán totalmente condicionados a esa primera fase de entrenamiento y de la representación correspondiente así como de la ponderación anteriormente citada dependerá la efectividad del sistema.

Se puede por tanto, decir que la representación de los documentos ha de ser la misma en la clasificación que en el entrenamiento, ya que será de esta manera la forma en la que un documento podrá verse a qué otros documentos se parece.

En próximos apartados se explicarán con más detalle las técnicas de representación y modelos de clasificación.

2.4.3.2 TIPOS DE CLASIFICADORES AUTOMÁTICOS

Los clasificadores automáticos de documentos se clasifican dependiendo de sus características y de ciertos criterios, según se muestra en la siguiente figura.

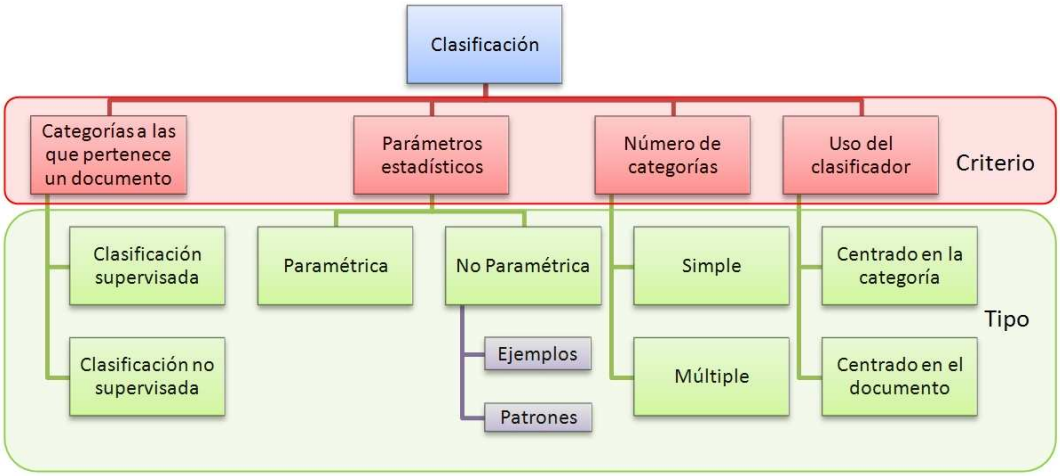


Ilustración 9 - Tipos de Clasificadores

2.4.3.2.1 CLASIFICACIÓN SUPERVISADA Y NO SUPERVISADA

Centrándose en la primera clasificación mostrada en el esquema anterior (Ilustración 9) y siguiendo el criterio de determinación de categorías a las que puede pertenecer un documento, se divide en dos tipos:

- **Clasificación Supervisada:** Se parte de una serie de clases o categorías conceptuales diseñadas de antemano. Asigna a cada documento una categoría. Este tipo de clasificación requiere de un corpus de entrenamiento previamente elaborado manualmente.

  Estos tipos de clasificadores pretenden elaborar un patrón representativo para cada una de las categorías entrenadas para después aplicar una función que permita calcular la similitud entre el documento y cada uno de los patrones adquiridos.

- **Clasificación no supervisada:** En este tipo de clasificadores no existe un corpus de entrenamiento previo, por el contrario, los documentos se clasifican en función de su contenido de manera automática. Este tipo de clasificación se suele conocer como *clustering*.

2.4.3.2.2 CLASIFICACIÓN PARAMÉTRICA Y NO PARAMÉTRICA

Si las técnicas de clasificación se basan en parámetros estadísticos como la media, la varianza u otros, o no, los clasificadores pueden diferenciarse en dos tipos:

- **Clasificación paramétrica:** En el entrenamiento de un clasificador paramétrico se emplea el *set* de entrenamiento para estimar o aprender los parámetros estadísticos del modelo.

  El *set* de test que contiene documentos a clasificar se emplea para determinar la capacidad de generalización del clasificador [Turner, K & Thost, 1995].
- **Clasificación no paramétrica:**
  - **Basada en patrones:** Se obtiene una descripción de cada categoría en términos de un patrón, típicamente en forma de vector de términos. La similitud de los documentos se realiza en función de las similitudes entre cada documento y los distintos patrones que representan las categorías [Bacan, Pandzic, & Guija, 2005]. Un ejemplo de este tipo es el clasificador Rocchio.
  - **Basada en ejemplos:** Los documentos se clasifican según la similitud que presentan con ejemplos del conjunto de entrenamiento. Un clasificador típico es el vecino más cercano (KNN, K-Nearest Neighbor) [Sebastiani, 2002].

2.4.3.2.3 CLASIFICACIÓN MÚLTIPLE Y SIMPLE

Un clasificador se puede dividir en dos tipos (múltiple y simple) en función del número de categorías en las que se puede clasificar un documento.

- **Simple:** Cada documento se clasifica en una única categoría. Las categorías no se solapan, es decir, un documento clasificado en una categoría A, nunca podrá ser clasificado en la categoría B. Un caso especial de este tipo de clasificación es la clasificación binaria, donde los documentos pertenecen a la categoría c_i o a su complementaria $\overline{c_i}$.
- **Múltiple:** Cada documento a clasificar puede pertenecer a un conjunto de categorías.

2.4.3.2.4 CLASIFICACIÓN CENTRADA EN LA CATEGORÍA Y EN EL DOCUMENTO

Existen dos formas de utilizar un clasificador automático, teniendo en cuenta el hecho de que el conjunto de categorías C o el conjunto de documentos D pueden que no se encuentren disponibles de forma completa desde el comienzo [Sebastiani, 2002].

- **Clasificación centrada en la Categoría:** Dado un documento, consiste en encontrar todas las categorías dentro de las cuales puede ser clasificado.

- **Clasificación Centrada en el Documento**: Dada una categoría, encontrar todos los documentos que pueden ser clasificados en dicha categoría.

2.4.3.3 TÉCNICAS Y ALGORITMOS DE CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

2.4.3.3.1 EL MODELO VECTORIAL

Es quizás el modelo más sencillo y es el mejor caso para explicar a grandes rasgos cómo funcionan en general los algoritmos de clasificación.

Este modelo intenta recoger la relación de cada documento D_i de una colección de N documentos, con el conjunto de las m características de la colección. Un documento puede expresarse como el vector que expresa la relación del documento con cada una de esas características.

$$D_i \rightarrow \vec{d} = (C_1, C_2, \dots, C_m)$$

Ecuación 1 - Modelo vectorial. Representación vectorial de un documento.

Donde C_{ik} es un valor numérico que expresa en qué grado el documento D_i posee la característica k, llamado peso.

Una vez seleccionado el conjunto de términos caracterizadores de la colección de documentos, es necesario calcular el valor de cada elemento del vector en el documento. El caso más simple es utilizar una aproximación binaria, de tal manera que si en el documento D_i aparece el término k, el valor C_{ik} sería 1, y en caso contrario sería 0.

La capacidad de representación de un término para un documento dado, se puede calcular hallando el número de veces que éste aparece en dicho documento (frecuencia del término en el documento, *term frecuency* - tf). Si la frecuencia de un término es extremadamente alta en el conjunto de documentos, se optará entonces por eliminarla. Por lo tanto, la capacidad de recuperación de un término es inversamente proporcional a su frecuencia en la colección de documentos ($inverse\ document\ frecuency - idf$).

$$w_j = tf_j \cdot idf_j$$

Ecuación 2 - Modelo Vectorial. Peso de un elemento

Así pues, el mecanismo de obtención de pesos también se aplica a las consultas, para de esta manera, poder disponer de representaciones vectoriales homogéneas de consultas y documentos, que posibiliten obtener el grado de similitud entre ambos documentos, representados como vectores de un espacio multidimensional.

El modo más simple de obtener una similitud entre una consulta y un documento, utilizando el modelo vectorial, es realizar el producto escalar de los vectores que lo representan [Venegas, 2007].

$$simil(Q_i, D_j) = \frac{\sum_{j=1}^{m} p_j \cdot d_{ij}}{\sqrt{\sum_{j=1}^{m} {p_j}^2 \cdot \sum_{j=1}^{m} {p_{ij}}^2}}$$

Ecuación 3 - Modelo Vectorial. Producto escalar

2.4.3.3.2 MODELO DE PROBABILÍSTICO DE BAYES

Como el propio nombre indica, se basa en la teoría probabilística, en especial en el teorema de Bayes:

Sea $C = \{A_1, A_2, \ldots, A_i, \ldots, A_n\}$ un conjunto de sucesos mutuamente excluyentes y exhaustivos, y tales que la probabilidad de cada uno de ellos es distinta de cero. Sea B un suceso cualquiera del que se conocen las probabilidades condicionales $P(B \mid A_i)$. Entonces, la probabilidad $P(A_i \mid B)$ viene dada por la expresión:

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{P(B)} = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{n} P(B|A_j)P(A_j)}$$

Ecuación 4 - Teorema de Bayes

Donde:

- $P(A_i)$ son las probabilidades a priori.
- $P(B \mid A_i)$ es la probabilidad de B en la hipótesis A_i.
- $P(A_i \mid B)$ son las probabilidades a posteriori.
- Esto se cumple $\forall i = 1 \ldots n$.

El algoritmo más conocido y a su vez el más simple es el denominado Naïve Bayes [Figuerola y otros. 2004] que, como es obvio, estima la probabilidad de que un documento pertenezca a una categoría. La pertenencia a la categoría depende de la posesión de ciertas características de las que se conoce la probabilidad de que aparezcan en los documentos que pertenecen a la categoría en cuestión. Las características que se comentan, son los términos de los documentos y tanto su probabilidad de aparición en general como en una categoría concreta, pueden observarse a partir de los datos de entrenamiento.

Un documento es representado mediante la expresión $\overrightarrow{d_j} = \{w_{ij}, \ldots, w_{|T|j}\}$ y una categoría viene dada por la expresión C_i . La probabilidad de que $\overrightarrow{d_j} \in C_i$ viene dada por la expresión [Sebastiani, 2002]:

$$P(C_i|\overrightarrow{d_j}) = \frac{P(C_i)P(\overrightarrow{d_j}|C_i)}{P(\overrightarrow{d_j})}$$

Ecuación 5 - Método probabilístico de Bayes. Probabilidad de que un documento pertenezca a una categoría

Donde $P(\overrightarrow{d_j})$ y $P(C_i)$ equivalen a la probabilidad de que un documento elegido al azar tenga como su representación el vector $\overrightarrow{d_j} \in C_i$.

La estimación de $P(\overrightarrow{d_j}|C_i)$ acarrea ciertos problemas ya que el número de posibles vectores $\overrightarrow{d_j}$ es más que elevado. Dicha probabilidad, Naïve Bayes, se calcula haciendo la suposición de que dos coordenadas cualesquiera son variables aleatorias estadísticamente independientes y viene representada con la siguiente ecuación:

$$P(\overrightarrow{d_j}|C_i) = \prod_{k=1}^{|T|} P(w_{kj}|C_i)$$

Ecuación 6 - Método probabilístico de Bayes. Probabilidad de que una categoría pertenezca a cierto documento.

Existen problemas cuando las colecciones de datos para el entrenamiento son pequeñas ya que pueden producirse errores al estimar las probabilidades. Estos problemas implican la necesidad de las llamadas técnicas de suavizado para evitar distorsiones en la obtención de las probabilidades [Figuerola y otros, 2004].

2.4.3.3.3 ALGORITMO DE ROCCHIO

Este algoritmo [Figuerola y otros, 2004] se aplica a la realimentación de consultas. Tras realizar y formular la primera consulta, el usuario examina los resultados del clasificador y determina cuáles resultan relevantes y cuáles no. Con estos datos, el sistema genera una nueva consulta basándose en los documentos que el usuario trató con anterioridad.

El algoritmo de Rocchio proporciona un sistema capaz de construir una nueva consulta recalculando los pesos de los términos de dicha consulta y aplicando un coeficiente a los pesos de la consulta inicial, otro a los documentos relevantes y otro al resto.

Por otra parte, es capaz de crear los patrones de cada una de las clases o categorías de documentos. Partiendo de una colección de entrenamiento, previamente categorizada, y aplicando el modelo vectorial, se pueden construir vectores patrón para cada una de las categorías, considerando como ejemplos positivos los documentos de entrenamiento de esa categoría, y como ejemplos negativos los de las demás categorías.

Dado el conjunto de entrenamiento T_r se construye el vector para la categoría $c_i = \{w_{1i}, w_{2i}, \dots, w_{ri}\}$ con la siguiente fórmula:

$$w_{ki} = \beta \cdot \sum_{(d_j \in POS_i)} \frac{w_{kj}}{|POS_i|} - \gamma \cdot \sum_{(d_j \in NEG_i)} \frac{w_{kj}}{|NEG_i|}$$

Ecuación 7 - Algoritmo de Rocchio. Construcción del vector para cada categoría

Donde:

- w_{kj} es el peso que tiene el término t_k en el documento d_j.
- $POS_i = \{d_j \in T_r | \widehat{\emptyset}(d_j, c_i) = T\}$.
- $NEG_i = \{d_j \in T_r | \widehat{\emptyset}(d_j, c_i) = F\}$.
- β y γ son parámetros para ajustar la importancia de los ejemplos negativos y positivos.

El vector de la categoría c_i representa el centroide de los documentos relevantes. El clasificador devuelve la proximidad que existe entre un documento test y el centroide de los documentos positivos y a su vez, la distancia al centroide de los documentos negativos.

Una vez realizado lo anterior, para clasificar un documento se simula la similitud entre el nuevo documento y cada uno de los patrones de las clases previamente clasificadas. Un esquema de este tipo de clasificador, podría ser el siguiente.

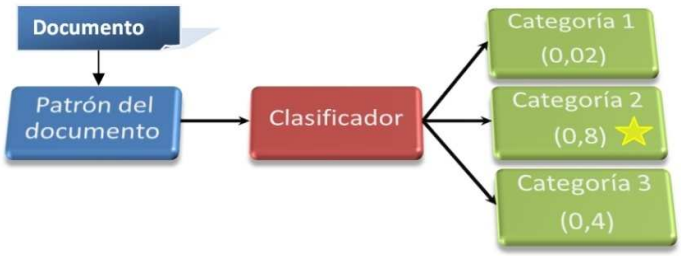


Ilustración 10 - Clasificador Rocchio

2.4.3.3.4 ALGORITMOS BASADOS EN EJEMPLOS

El aprendizaje basado en ejemplares o instancias, tiene como principio fundamental, el almacenamiento de ejemplos. La clasificación posterior se realiza por medio de una función que mide la proximidad entre el documento a clasificar y los ejemplos de la base de entrenamiento.

El algoritmo de $k - vecinos\ más\ próximos$ del inglés $KNN,\ k - Nearest\ Neighbor$ representa este tipo de algoritmos. Es un método no paramétrico ya que no se conoce ninguna suposición distribucional acerca de las variables predictoras. Los ejemplos son vectores multidimensionales, donde cada uno viene descrito por un conjunto en términos $|T|$ atributos $x_j = \{w_{1j}, \dots, w_{|T|j}\}$ y $|C|$ categorías son consideradas.

Para inferir la categoría de un documento, el algoritmo compara dicho documento con todos los ejemplos de entrenamiento y calcula la distancia entre ellos. Típicamente, la clase mayoritaria entre los k primeros ejemplos obtenidos, aquellos que son, por lo tanto, más similares al obtenido, es la categoría inferida por el sistema. Generalmente se usa la distancia Euclídea

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^{|c|} (x_{ik} - x_{jk})^2}$$

Ecuación 8 - Algoritmos basados en ejemplares. Distancia Euclídea

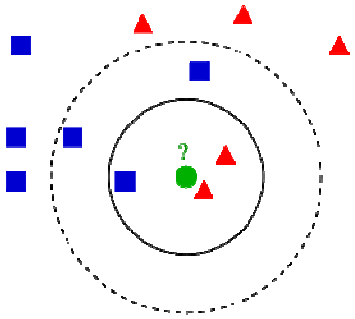


Ilustración 11 - Ejemplo de KNN

En la ilustración anterior se puede ver un ejemplo claro. En la fase de entrenamiento el sistema ha obtenido los valores de sus atributos (distribución en el plano) y los valores de sus clases (forma y color). El sistema pretende entonces, al clasificar, decidir de qué ejemplos está más cerca el documento (círculo verde).

2.4.3.3.5 ÁRBOLES DE DECISIÓN

Son quizás la forma más sencilla de representar el conocimiento, y es uno de los métodos de aprendizaje inductivo supervisado no paramétrico más utilizado [Cortijo Bon, 2000].

Un árbol de clasificación es una forma de representar el conocimiento obtenido en el proceso de aprendizaje inductivo. Es una partición recursiva del espacio representativo del conjunto de documentos de entrenamiento, es decir, una organización jerárquica del espacio de representación que puede modelarse mediante una estructura de tipo árbol. Cada nodo interior contiene una pregunta sobre un atributo concreto, con un hijo por cada respuesta, y cada nodo hoja se refiere a la decisión, es decir, a la clasificación.

La clasificación de patrones se realiza en base a una serie de preguntas sobre los valores de sus atributos, empezando por el nodo raíz, y siguiente el camino que indican las respuestas a las preguntas de nodos internos. Así hasta llegar a un nodo hoja, que contiene la clasificación.

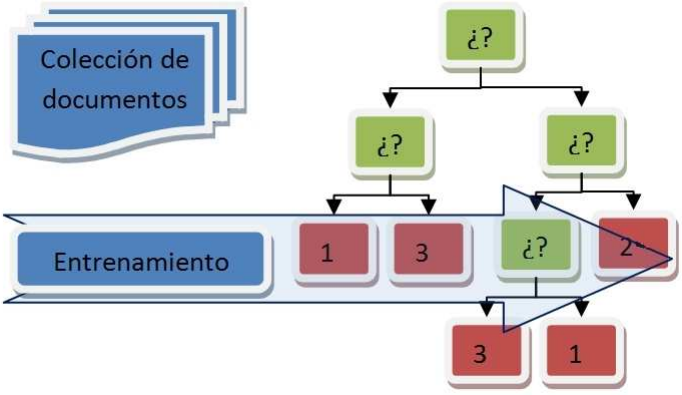


Ilustración 12 - Árboles de decisión. Entrenamiento

El aprendizaje o entrenamiento en un sistema de árboles de decisión consiste en la construcción del árbol a partir de un conjunto de prototipos.

Por otra parte, la clasificación consiste en el etiquetado de un patrón independientemente del conjunto de aprendizaje. Es por tanto una tarea de responder a las preguntas a asociadas a los nodos interiores utilizando los valores de los atributos del patrón del documento a clasificar. Este proceso se repite hasta alcanzar un nodo hoja.

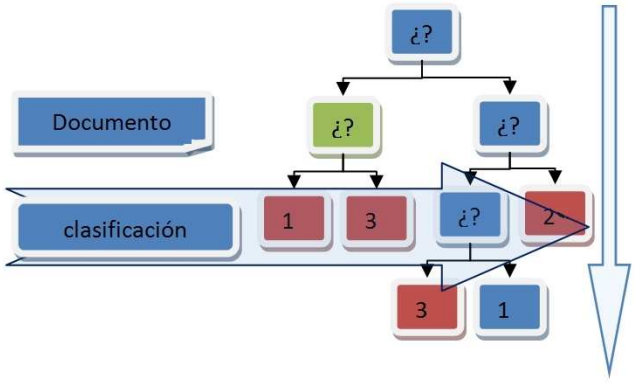


Ilustración 13 - Árboles de decisión. Clasificación

2.4.3.3.6 MÁQUINA DE VECTORES DE SOPORTE

Son una técnica de gran utilidad cuando se quiere construir un clasificador utilizando ejemplos. La máquina de vectores soporte (SVM, *Support Vector Machines*) se basa en el principio de inducción de Minimización del Riesgo Estructural (SRM, *Structural Risk Minimization*) como proceso de inferencia [Resendiz, 2006].

El procedimiento se basa en encontrar una hipótesis h para la cual, a partir de la obtención de una cota sobre el riesgo esperado $R(h)$ (tasa de error medio sobre el conjunto de test) se concluye que, para asegurar su minimización, fijado el conjunto de entrenamiento, es necesario minimizar conjuntamente el riesgo empírico $R_{emp}(h)$ (tasa de error media sobre el conjunto de entrenamiento) y la VC (Vapnik Chervonenkis) dimensión del espacio de hipótesis [Cabello Pardos, 2004].

El riesgo empírico viene dado por la expresión [David & Lerner, 2004]:

$$R_{emp}(h) = \frac{1}{2n}\sum_{i=1}^{n}|y_i - f(x_i, h)|$$

Ecuación 9 - Máquinas de vectores soporte. Riesgo Empírico

Donde:

- n es el tamaño del documento de entrenamiento.
- $f(x_i, h)$ es la salida del clasificador por un vector de entrenamiento x_i.
- $y_i \in \{-1,1\}$.

El riesgo esperado por un vector del conjunto de test x viene dado por:

$$R(h) = \int \frac{1}{2}|y - f(x,h)| \cdot dP(x,y)$$

Ecuación 10 - Máquina de vectores soporte. Riesgo esperado para un vector del conjunto de entrenamiento

No obstante, en esta fórmula, $dP(x,y)$ es desconocido. En 1995, Vapnik demostró que, con una probabilidad de $1-\mu$ con $0 \leq \mu \leq 1$, una cota superior de riesgo esperado se puede obtener con la siguiente ecuación [Vapnik, 1995]:

$$R(h) \leq R_{emp}(H) + \sqrt{\frac{d(\ln\frac{2n}{d}+1) - \ln\frac{\mu}{4}}{n}}$$

Ecuación 11 - Máquina de vectores soporte. Riesgo esperado de Vapnik

Donde d es un entero no negativo como la dimensión VC

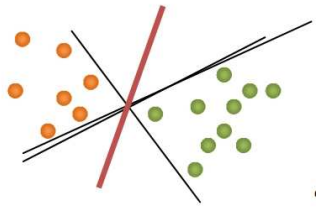


Ilustración 14 - Hiperplano óptimo

El principio básico de las máquinas de vectores soporte es la obtención y selección de la hipótesis que da lugar al margen mayor de separación entre clases, llamado hiperplano de separación óptimo. En la anterior figura se puede observar cómo ciertos clasificadores lineales (líneas negras) pueden separar los datos, pero únicamente un hiperplano (línea roja) maximiza la distancia entre él y el punto más cercano de cada clase.

2.4.3.3.7 REDES NEURONALES

Las redes neuronales han sido propuestas en numerosas ocasiones como instrumentos útiles para la RI y también para la clasificación automática. De una manera genérica, una de las principales aplicaciones de las redes neuronales es el reconocimiento de patrones. Por tanto, no es de extrañar que se hayan aplicado a problemas de categorización de documentos [Figuerola y otros. 2004].

Una red neuronal consta de varias capas de unidades de procesamiento o neuronas interconectadas. En el ámbito que aquí ocupa, la capa de entrada recibe términos, mientras que las unidades o neuronas de la capa de salida mapean clases o categorías.

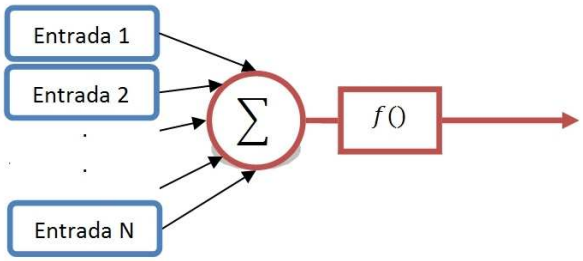


Ilustración 15 - Neurona

Una neurona es un dispositivo sencillo formado por una serie de entradas y una única salida [Cabello Pardos, 2004]. Cada neurona acepta como entrada las salidas procedentes de otras neuronas, siendo la entrada efectiva a la neurona la suma ponderada de las entradas reales a dicha neurona. Cada neurona se caracteriza por su estado de activación, que típicamente es un valor que oscila entre 0 y 1. Si el estado de activación de una neurona es 0, la neurona no está activada; mientras que cualquier valor distinto de 0 corresponde a una neurona activa. La salida de la neurona es el estado de activación. Cada neurona realiza una tarea sencilla: recibe la información de entrada de las neuronas o del exterior y la usa para calcular una señal de salida que se propaga a otras unidades.

Las interconexiones tienen pesos, es decir, un coeficiente que expresa la mayor o menor fuerza de la conexión. Es posible entrenar una red para que, dada una entrada determinada (los términos de un documento), produzca la salida deseada (la clase que corresponde a ese documento). El proceso de entrenamiento consta de un ajuste de los pesos de las interconexiones, a fin de que la salida sea la deseada.

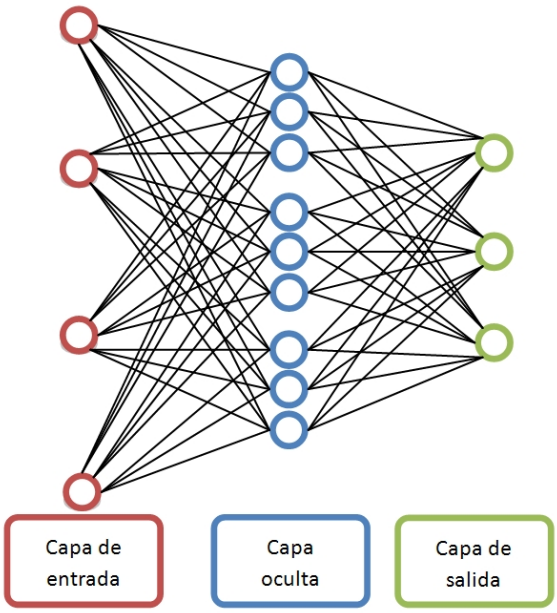


Ilustración 16 - Perceptrón multicapa

En general, las neuronas se organizan en capas. Dependiendo de su función en la red, se distinguen tres tipos de unidades: las unidades cuya activación son los datos de entrada del problema (unidades de entrada); las unidades cuya salida es la salida del problema (unidades de salida); y el resto de unidades, llamadas unidades ocultas (ya que no son "visibles" desde el exterior). Esta disposición de la red da lugar al llamado perceptrón multicapa.

2.4.4 EVALUACIÓN DE UN SISTEMA DE CLASIFICACIÓN AUTOMÁTICA DE VÍDEOS

Los sistemas de clasificación automática de vídeos deben, como todo sistema, evaluarse en global, su precisión, velocidad, exhaustividad, efectividad, entre otros, deben ser analizadas en conjunto, pero estos sistemas tienen claramente dos partes bien definidas, el reconocimiento automático del habla y la clasificación automática de documentos.

El resultado de las pruebas de la clasificación automática de documentos, será en realidad el resultado final del sistema ya que, las clasificaciones o categorías que devuelva, serán las que condicionen las características del mismo. Cabe destacar que los errores arrastrados del reconocimiento automático del habla afectan directamente al sistema.

Cada una de las partes del sistema son en realidad otros sistemas. El siguiente esquema resume las características importantes de un sistema, los errores arrastrados y cómo afectan en el sistema global.

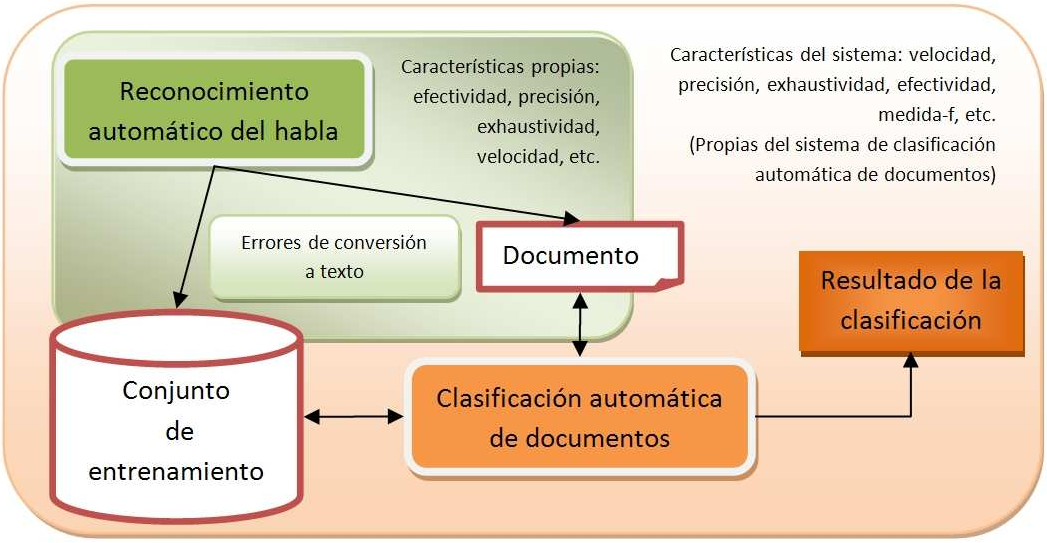


Ilustración 17 - Errores y características en la clasificación automática de vídeos

Como se puede observar, el error cometido por el reconocimiento automático del habla condiciona el sistema ya que, de la precisión en el texto obtenido partirá el clasificador de documentos.

Es fácil observar que un cambio de la palabra *ola* por *hola*, que suenan igual, implica un error en la búsqueda de documentos o categorías.

El siguiente ejemplo muestra un posible error arrastrado de la transcripción y el problema que acarrea al sistema en global.

> ***Texto original del vídeo:*** *Las carreras de coches crean expectación entre los más jóvenes, motor, adrenalina y sexo son los ingredientes de estas jornadas.*
>
> ***Texto tras la transcripción:*** *Las camareras de noches crean expectación entre los más jóvenes, motor, … y sexo son los ingredientes de estas …*

Ejemplo 9 - Ejemplo de error en la transcripción automática de un audio.

La confusión de dos palabras y la omisión de otras dos, hace que tanto el significado de la frase como la importancia de algunos términos desaparezcan o cambien. El texto dado por el sistema de reconocimiento del habla será la base para la clasificación del texto, por lo que, dicho texto desde un principio ya será erróneo y seguramente sea clasificado erróneamente.

Los clasificadores automáticos necesitan de técnicas y algoritmos que permitan obtener qué documentos se parecen entre sí o qué documentos pertenecen a qué categoría.

2.4.4.1 EVALUACIÓN GENERAL DE LOS SISTEMAS DE CLASIFICACIÓN

Para saber qué técnica o qué algoritmo es mejor es necesario hacer un estudio de los mismos teniendo en cuenta las siguientes características:

- **Precisión (en inglés *precisión*):** Representa el nivel de confianza del clasificador y viene establecido por la siguiente fórmula:

$$precision = \frac{tp}{tp + fp}$$

Ecuación 12 - Precisión

Un sistema puede acertar siempre la categoría de cada documento porque clasifica pocos documentos, es decir, es un sistema muy preciso pero poco exhaustivo.

Donde:

- TP_i (*True positives*, verdaderos positivos): Representan el total de documentos que han sido correctamente clasificados en la categoría c_i
- FP_i (*False positives*, falsos positivos): número de documentos clasificados en la categoría c_i siendo esto erróneo.
- FN_i (*False negatives*, falsos negativos): Numero de documentos de c_i que no han sido clasificados como tal.
- TN_i (True negatives, verdaderos negativos): Aquellos documentos que no pertenecen a c_i y no han sido asignados en ella.

- **Exhaustividad (en inglés *recall*, cobertura):** Representa la cobertura del clasificador, es decir, la cantidad de documentos que clasifica frente a los no clasificados y clasificados. Un sistema puede clasificar todos los documentos en una categoría, aunque lo haga mal, teniendo pues una exhaustividad alta pero una precisión baja.

$$exhaustividad = \frac{tp}{tp + fn}$$

Ecuación 13 - Exhaustividad (Recall)

- **Medida-F (en inglés F-measure):** Realiza una media entre la precisión y la exhaustividad para poder obtener un resumen de la eficacia. Dependiendo del valor de Beta se asigna más peso a la precisión ($\beta < 0.5$) o a la cobertura ($\beta > 0.5$). Habitualmente $\beta = 0.5$, valorandose con la misma importancia a la precisión y al *recall*.

$$medidaF = \frac{(1 + \beta^2) * precisión * cobertura}{(\beta^2 * precisión) + cobertura}$$

Ecuación 14 - Medida-F

- **Lift:** Representa la capacidad de un sistema de predecir la categoría frente a una elección al azar, es decir, si un clasificador tiene una precisión del 60% y la probabilidad de acertar una categoría al azar es del 10% (1 entre el número de categorías, en este caso 10), el sistema por lo tanto, tendrá una precisión 6 veces superior a la elección de una al azar, lo que se puede considerar algo bastante aceptable.

$$lift = \frac{\frac{tp}{tp + fp}}{\frac{1}{num\ de\ categorías}} = \frac{precisión\ del\ sistema}{probabilidad\ de\ acertar\ al\ azar}$$

Ecuación 15 - Lift

- **Velocidad:** La velocidad de ejecución hoy día es una de los factores más importantes. En sistemas de *clustering*, por ejemplo, donde se tienen que clasificar cientos de noticias al segundo, de nada sirve la precisión o la exhaustividad si el sistema es lento.
- **Claridad:** Las reglas que permiten al sistema realizar la clasificación deben ser simples y sencillas y sobre todo creíbles por el usuario.
- **Tiempo de aprendizaje:** El tiempo de aprendizaje debe ser lo más rápido posible ya que de esta manera un clasificador podrá adaptarse a nuevas categorías o añadir documentos a las categorías, siendo así un clasificador dinámico.

2.4.4.2 EVALUACIÓN DE LOS SISTEMAS DE RECONOCIMIENTO AUTOMÁTICO DEL HABLA

Aunque no es un tema de estudio en el presente proyecto, es quizás importante conocer cómo se evalúan estos sistemas.

Para evaluar la calidad de transcripción es necesario comparar la salida del sistema automático con la original y calcular ciertas medidas de evaluación. La tasa de error de palabras, WER (del inglés *Word Error Rate*), es una medida comúnmente usada en la evaluación de sistemas de reconocimiento del habla y traducción automática [Gallo, San-Segundo, 2008].

Esta medida calcula el número de borrados, inserciones y sustituciones de palabras cuando se comparan frases. Se basa en la distancia de edición o de Levensthein, es decir, el número mínimo de operaciones para transformar una cadena de caracteres en otra. En sistemas de reconocimiento del habla se calcula el WER entre la frase generada por el sistema y una frase de referencia correcta.

$$WER = \frac{S + B + I}{N}$$

Ecuación 16 - Tasa de Error de palabra (Word Error Rate)

Donde:

- S es el número de sustituciones.
- B es el número de borrados.
- I es el número de inserciones.
- N es el número de palabras que tiene la frase referenciada.

Existen otras maneras menos estrictas de evaluar un sistema de reconocimiento del habla. El BLEU (*Bilingual Evaluation Understudy*) [Papineni, Roukos, 2002] es un método de evaluación de la calidad de traducciones realizadas por sistemas automáticos. Una traducción tiene mayor calidad cuanto más similar es con respecto a otra de referencia.

BLEU se calcula normalmente a nivel de frases y halla la precisión en $n-gramas$ entre la traducción del sistema y la de referencia. Estas medidas surgen con el objetivo de encontrar medidas automáticas que corresponden con la evaluación que un experto haría de la traducción.

Antes de calcular el BLEU es necesario comentar ciertas medidas en las que se basa, como la precisión de $n-gramas$ entre dos frases, que se calcula con la siguiente fórmula:

$$p = \frac{n-gramas\ Comunes}{n-gramas\ Candidata}$$

Ecuación 17 - Precisión en n-gramas (BLEU)

El siguiente ejemplo servirá para entender mejor lo comentado.

Referencia: El reconocimiento automático y el habla
Candidata1: El reconocedor automático y el habla
Candidata2: El reconocedor autómata y el hablar

Candidata3: El el el el el el
Candidata4: El el

Ejemplo 10 - Precisión en n-gramas (BLEU)

Cuando una frase es de menor tamaño que la correcta (candidata4) se ve reflejado en la precisión modificada anterior. La candidata4 tiene una precisión de $2/2$, como no refleja la similitud entre ambas frases existe un penalizador:

$$PB = \begin{cases} 1 \; si \; c > r \\ e^{1-\frac{r}{c}} \end{cases}$$

Ecuación 18- Penalización a precisión en n-gramas (BLEU)

Donde:

- c es la longitud de la frase candidata
- r es la longitud de la frase referida

Entonces, para calcular el BLEU se hará de la siguiente manera:

$$BLEU = PB \cdot exp\left(\sum_{n=1}^{N} \omega_n \log P_n\right)$$

Ecuación 19 - Bilingual Evaluation Understudy (BLEU)

Donde:

- ω_n es el peso de cada $n - grama$ y viene dado por la expresión $\omega_n = \frac{1}{N}$

La precisión en $1 - gramas$ es $4/6$ para la candidata1 y de $2/6$ para la candidata2, pero es importante destacar que la candidata3 tendría una precisión de $6/6$, es decir, una precisión de 100%. Es por esta razón por lo que hay que tener en cuenta el número máximo de ocurrencias de un $n - grama$ en la frase de referencia, siendo éste el límite a la hora de contabilizar las apariciones en la frase candidata. Por lo tanto, la precisión modificada quedaría entonces, en la candidata3, de $2/6$.

2.4.4.3 EVALUACIÓN DE LOS SISTEMAS DE CLASIFICACIÓN AUTOMÁTICA DE TEXTOS

La evaluación de los sistemas de clasificación automática de textos se realiza de forma experimental ya que de no ser así, para poder hacerlo analíticamente se necesitaría una especificación formal del problema a resolver. Es por ello que la evaluación de los clasificadores es experimental, evaluando la capacidad de dar los resultados correctos, es decir, su efectividad.

Para evaluar un sistema de clasificación de este tipo, los procesos y medidas que se van a realizar en el experimento deben ser los mismos. Las pruebas serán entonces las siguientes: Dada una colección de datos, llamado corpus, una parte de la misma es considerada como conjunto de entrenamiento y el resto como conjunto de test. De esta

manera, se pretende que el modelo aprenda del conjunto de entrenamiento e infiera las categorías para los ejemplos del conjunto de test.

Para comprobar la efectividad de un sistema de clasificación automática es necesario emplear las medidas típicas de cobertura (exhaustividad o *recall*) y precisión que ya se comentaron.Una medida característica de estos sistemas es la precisión con respecto a la categoría así como la cobertura. Se define $\breve{\phi}(d_x, c_i)$ como el evento de que d_x sea clasificado en c_i y $\phi(d_x, c_i)$ como el evento de que d_x pertenece a c_i.

Por tanto, la precisión con respecto a la categoría se define como la probabilidad condicional de que un documento d_x pertenezca a la categoría c_i siendo entonces verdadero (T):

$$P\left(\breve{\phi}(d_x, c_i) = T | \phi(d_x, c_i) = T\right)$$

Ecuación 20 - Precisión con respecto a la categoría

Por otra parte, la cobertura con respecto a la categoría se define como la probabilidad de que, si un documento d_x debe ser clasificado bajo la categoría c_i esta decisión es tomada, es decir, la probabilidad de que un documento d_x sea clasificado en su categoría correspondiente.

$$P\left(\phi(d_x, c_i) = T | \breve{\phi}(d_x, c_i) = T\right)$$

Ecuación 21 - Cobertura con respecto a la categoría

La tabla de contingencia de una categoría c_i muestra lo explicado anteriormente y se resume en la siguiente tabla.

Categoría c_i		Criterio del experto	
		SÍ	NO
Criterio del clasificador	SÍ	TP_i	FP_i
	NO	FN_i	TN_i

Tabla 1 - Tabla de contingencia de una categoría

La precisión y la cobertura referentes a una categoría c_i siguen la fórmula ya comentada:

$$Precisión\ en\ la\ categoría\ c_i = \pi_i = \frac{TP_i}{TP_i + FP_i}$$

$$Cobertura\ en\ la\ categoría\ c_i = p_i = \frac{TP_i}{TP_i + FN_i}$$

Ecuación 22 - Precisión y Cobertura de una categoría

Existe una medida que resumen las dos anteriores comparándolas, es la llamada *medida-F* (Ecuación 14 - Medida-F).

2.4.4.3.1 EVALUACIÓN DE CATEGORÍAS

Típicamente, para evaluar un sistema de clasificación automática, se realiza un análisis de cada categoría o clase del sistema. Para ello se obtiene la siguiente tabla.

Real / Predicha	Categoría 1	Categoría 2	…	Categoría N	Instancias reales
Categoría 1	Cr_1s_1	Cr_1s_2	…	Cr_1s_n	$\sum_{j=1}^{n} Cr_1s_j$
Categoría 2	Cr_2s_1	$C_{r2}s_2$	…	Cr_2s_n	$\sum_{j=1}^{n} Cr_2s_j$
…	…	…	…	…	…
Categoría N	Cr_ns_1	Cr_ns_2	…	Cr_ns_n	$\sum_{j=1}^{n} Cr_ns_j$

Tabla 2 - Matriz de confusión

Esta tabla representa la matriz de confusión entre clases del sistema, es decir, la manera en que el sistema se confunde entre clases.

Cada fila representa las instancias de la clase total, es decir, las instancias reales, mientras que cada columna representa las predicciones del sistema. Por lo tanto Cr_ns_j (Clase real, Clase sistema) representa el número de objetos de la clase n clasificados en la clase j.

Para observar las características del sistema en global se van a describir dos métodos, el micro-averaging y el macro-averaging, donde, respectivamente, se pretende dar igual peso a cada documento o igual peso a cada categoría.

2.4.4.3.2 MICRO-AVERAGING

Como se ha comentado, dando el mismo peso a cada documento, obtiene la precisión y la cobertura del sistema. Se resumen en la siguiente tabla y ecuaciones.

Categoría $C = \{c_1, \dots, c_{\|c\|}\}$		Criterio del experto	
		SÍ	NO
Criterio del clasificador	SÍ	$TP = \sum_{i=1}^{\|c\|} TP_i$	$FP = \sum_{i=1}^{\|c\|} FP_i$
	NO	$FN = \sum_{i=1}^{\|c\|} FN_i$	$TN = \sum_{i=1}^{\|c\|} TN_i$

Tabla 3 - Tabla de contingencia para micro-averaging

$$\pi^{\mu} = \frac{TP}{TP + FP}$$

$$p^{\mu} = \frac{TP}{TP + FN}$$

Ecuación 23 - Cobertura y precisión para micro-averaging

Donde $TP, FP\ y\ FN$ vienen establecidas por la tabla anterior.

2.4.4.3.3 MACRO-AVERAGING

En este caso, tanto la precisión como la cobertura se evalúan primeramente de forma local para cada categoría, después se hace una media con los resultados obtenidos.

$$\pi^{M} = \frac{\sum_{i=1}^{|c|} \pi_i}{|c|}$$

$$p^{M} = \frac{\sum_{i=1}^{|c|} p_i}{|c|}$$

Ecuación 24 - Cobertura y precisión en Macro-averaging

2.4.4.3.4 EVALUACIÓN CON N RESULTADOS

Cuando un sistema de clasificación automática de documentos posee más de dos categorías y quiere calcularse su precisión en un número n de resultados, por ejemplo la precisión dando 5 resultados, las medidas precisión, cobertura y medida-F se expresan de diferente manera a lo comentado hasta ahora.

Un documento será acertado si se acierta en alguno de los resultados dados, mientras que los fallos, solo se contarán si no se acierta en ninguno de los resultados dados. Este tipo de evaluación permite hacerse una idea de manera bastante global, tanto de la precisión como de la cobertura, tendiendo ambas a ser bastante altas y aumentando a medida que se van dando más resultados. La tabla que se muestra a continuación resume lo anterior.

Documentos	A1	…	An
Documento 1	A_{11}	…	A_{n1}
…	…	…	…
Documento j	A_{j1}	…	A_{jn}
No Clasificados	$TotalNo_1$		$TotalNo_2$
Total	$TotalAc_1$	…	$TotalAc_n$

Tabla 4 – Ejemplo de documentos clasificados - Precisión en n categorías

Dada la anterior tabla que representa un conjunto de documentos clasificados automáticamente, se dice:

- La precisión del sistema es la suma de los aciertos en 1 o 2 o … o n dividido entre el conjunto de documentos clasificados, viene dada por la siguiente expresión:

$$precisión_n = \frac{TotalAc_n}{TotalAc_n + TotalF_n}$$

Ecuación 25 - Precisión en n categorías

  Donde:
  - $TotalF_n$ es el total de fallos
    $$TotalF_n = DocumentosTotales - TotalNo_n - TotalAc_n$$
  - A_{jn} es acierto si $A_{j1}\ es\ acierto$ o $A_{j2}\ es\ acierto\ o$ … o $A_{jn}\ es\ acierto$
  - Un documento es no clasificado si no ha sido clasificado desde $1\ hasta\ n$
- La cobertura se calcula siguiendo la siguiente fórmula:

$$cobertura_n = \frac{TotalAc_n}{TotalAc_n + TotalNo_n}$$

Ecuación 26 - Cobertura en n categorías

2.5 ARQUITECTURA DE UN SISTEMA DE CLASIFICACIÓN AUTOMÁTICA DE VÍDEOS

La clasificación automática de vídeos o documentos audiovisuales es un problema poco abordado en la actualidad. Existen algunos eventos puntuales descritos en el estado del arte, como CLEF, y proyectos concretos que se dedican y trabajan en este tema, pero por el momento no existe una investigación madura en el asunto.

Este apartado pretende realizar una visión global sobre el diseño, arquitectura y procesos típicos de estos sistemas de clasificación automática de vídeos.

Un clasificador automático de vídeos, como ya ha sido comentado, consta de varios procesos. Se necesita por un lado un sistema capaz de reconocer el audio y por otro lado un sistema capaz de encontrar documentos similares a uno dado [Ide, 1999] [Jeong, 2002].

Insistiendo en lo ya descrito en otros apartados, un sistema de clasificación de textos requiere de una base de entrenamiento con la que poder clasificar, por esta razón es necesario distinguir, entonces, dos procesos claramente independientes, por un lado la clasificación automática y por otro el entrenamiento.

En primer lugar, un clasificador automático de vídeos necesita tener una base de datos de vídeos (corpus de entrenamiento y test) con las que poder entrenar su modelo de clasificación [Perea-Ortega, 2008]. Estos vídeos son procesados para extraer su audio y poder así obtener su transcripción.

Una vez obtenidas las transcripciones de los documentos audiovisuales es necesario pasarlas a un formato adecuado, para ser procesados tanto en el entrenamiento como en la clasificación.

La fase de entrenamiento necesita un gran corpus de documentos previamente clasificados, la siguiente ilustración muestra cómo un sistema de clasificación automático de vídeos procesa su corpus de entrenamiento previamente proporcionado. El sistema extrae el audio de cada uno de los vídeos de la base de datos que tiene, procesa sus audios y extrae sus conversaciones, para posteriormente convertir cada vídeo en un documento en texto plano. Una vez completada esta operación, es el clasificador automático de documentos el que se encarga de crear su estructura de datos, indexando cada uno de los documentos obtenidos previamente y entrenando así su algoritmo de aprendizaje. Como se verá en el siguiente capítulo, cada uno de los documentos del corpus es procesado con diversas técnicas.

Es importante observar esto desde un punto de vista de procesado del corpus que se le ha proporcionado al sistema y no como un sistema de RI, es decir, el sistema no extrae los audios, simplemente procesa aquellos vídeos previamente clasificados por un experto.

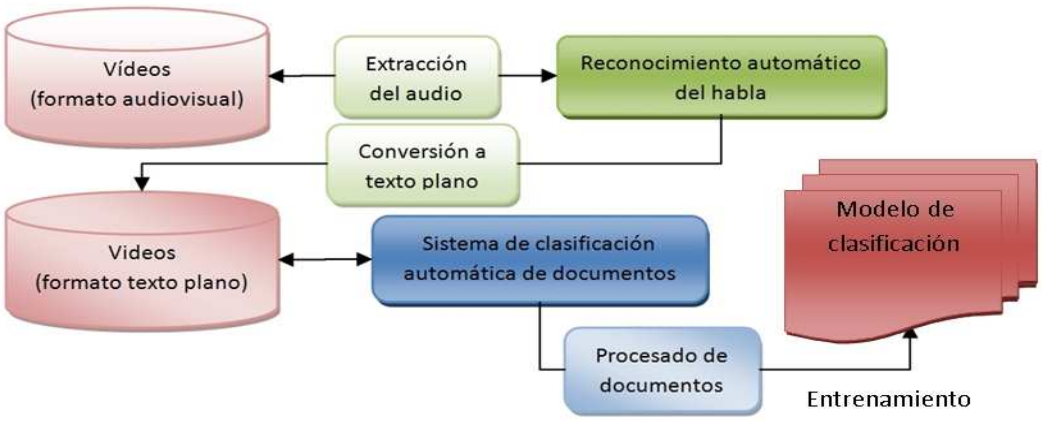


Ilustración 18 - Arquitectura de un sistema de clasificación de vídeos. Entrenamiento

La clasificación de un vídeo requiere prácticamente de los mismos procesos anteriormente comentados, de hecho, no es más que la comparación de un documento nuevo con la lista de documentos obtenidos en el proceso anterior.

La arquitectura mostrada en la Ilustración 18 y en la Ilustración 19, junto con las explicaciones dadas en otros capítulos, deja bastante claro que el gran peso del sistema reside en la clasificación automática de documentos, ya que, el reconocimiento del habla es como una caja negra que transforma los documentos audiovisuales en documentos en texto plano. Es por ello que a partir de este momento el documento se va a centrar en la clasificación automática de textos.

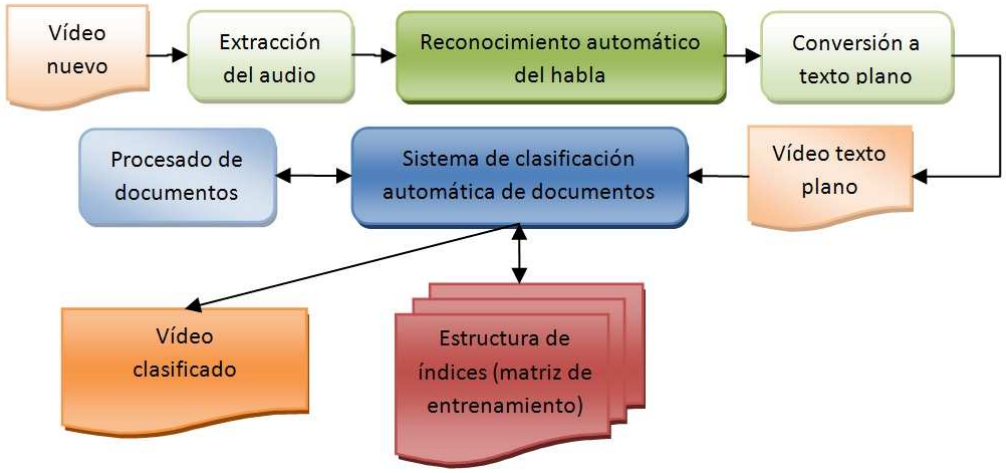


Ilustración 19 - Arquitectura de un sistema de clasificación de vídeos. Clasificación

2.5.1 FASES DE UN CLASIFICADOR AUTOMÁTICO DE DOCUMENTOS

Un clasificador automático de documentos requiere de ciertas fases de procesamiento de documentos donde, como previamente se ha dicho, los documentos adquieren una representación, se reducen sus dimensiones y se procesan sus datos para adquirir una ponderación y una asignación de pesos [Villena, Lana, 2008] [Perea-Ortega, 2008].

Claramente, ciertas fases deben ser las mismas tanto para un documento que forme parte del corpus de entrenamiento como para un documento que va a ser clasificado. Se pueden distinguir las siguientes fases.

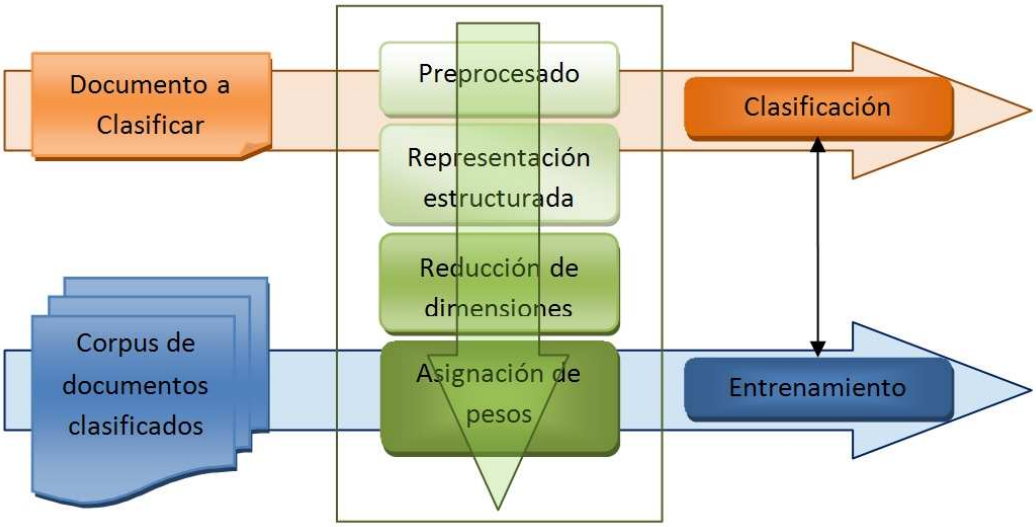


Ilustración 20 - Entrenamiento y clasificación automática.

- **Preprocesado y representación del documento:** Esta fase consiste en transformar cada documento en una representación más adecuada. Esta representación tiene que ser empleada por un clasificador.
- **Reducción de dimensiones:** el modelo que representa cada documento ha de contener únicamente aquellos que se consideran más importantes, despreciando así aquellos que no relevantes en el documento.
- **Asignación de pesos:** cada uno de los términos que posee cada documento ha de ponderarse dependiendo de la importancia que tiene en el documento.
- **Entrenamiento:** Esta fase consiste en la creación de un modelo dónde un conjunto de test será previamente clasificado a mano y representado y procesado cómo previamente se ha indicado.
- **Clasificación:** En función del conocimiento adquirido en la fase de entrenamiento en cuanto a cada una de las categorías, el clasificador asignará categorías a nuevos documentos.

Por lo tanto, un documento, en la fase de clasificación y en la fase de entrenamiento requiere de ciertos procesos comunes. La Ilustración 20 representa la manera en la que un documento se clasifica y la manera en la que un conjunto de documentos pasa a ser el conjunto de condiciones de pertenencia a una clase.

2.5.1.1 PREPROCESADO DEL DOCUMENTO Y REPRESENTACIÓN ESTRUCTURADA

La construcción de un clasificador automático de documentos presenta como punto de partida la obtención del conjunto de documentos o corpus. Puesto que los documentos no pueden ser interpretados directamente por el clasificador, la obtención de una representación compacta de cualquier documento d_i es necesaria tanto para el entrenamiento como para la fase de clasificación. A este proceso se le conoce como *indexación* y dependerá de la elección

que se considere como unidad mínima de información y las posibles reglas lingüísticas del idioma del documento.

Es común el uso, en este proceso, de herramientas del procesamiento del lenguaje natural. Se comentaron en el Capítulo 2 Capítulo 2 -ciertas técnicas que pretenden entender y procesar el lenguaje. Estas técnicas pueden ser usadas en el preprocesado de los textos. Las técnicas más usadas son las siguientes [Fernández, 2006]:

- **Palabras "vacías":**

  Los sistemas suelen evitar o eliminar las llamadas palabras vacías, que no son más que palabras que no aportan significado al documento.
- **Expansión por sinónimos:**

  Como es lógico un programa no puede esperar que el documento contenga la palabra justa con la que se tiene definida esa misma en el conjunto del corpus de entrenamiento.

  Algunos sistemas avanzados realizan una expansión por sinónimos en la consulta, como es el caso de los sistemas de respuesta automática, donde para intentar entender mejor la pregunta que se le ha realizado se utiliza la expansión de sinónimos que se ejecuta sobre las palabras que no son vacías, añadiendo los sinónimos de estas palabras a la consulta hará en el índice. Así hay más posibilidad de contestar adecuadamente a la pregunta, ya que en un buen programa que utilice análisis sintáctico podrá definir cada parte de la frase, y utilizar la acepción correspondiente en cada caso, ya que una palabra dependiendo del contexto puede tener significados totalmente distintos.
- **Lematización:**

  Lematizar consiste en la reducción de las diferentes formas flexivas de una palabra a la forma canónica, su lema, es decir, reagrupar las distintas inflexiones de un verbo en el infinitivo; el singular y el plural de un sustantivo en el singular; el masculino y el femenino de un adjetivo en el masculino, y lo mismo con adverbios, preposiciones, pronombres, etc.

  Con esto se consigue identificar familias de palabras para considerarlas como una sola. Es un dato muy importante para la búsqueda de información ya que se consigue hacer independiente la búsqueda de tiempos verbales y otros.

En cuanto a la representación estructurada del documento el modelo típico es el espacio de vectores, ya explicado en apartados anteriores (2.4.3.3.1 El modelo vectorial).

2.5.1.2 REDUCCIÓN DE DIMENSIONES

El vector de términos que representa al documento debe ser un vector de términos reducido y con la información estrictamente necesaria, eliminando aquellas palabras o términos que no son necesarios, es decir, almacenando únicamente aquellos que son más representativos del documento. Hay que seleccionar aquellas palabras clave que aportan significado y descartar aquellas otras que no contribuyen en la distinción entre documentos.

En gran medida de este proceso dependerá la efectividad del sistema y por otra parte se evita el sobreentrenamiento (*overfitting*) [Calvo, 2000].

La elección de las palabras clave se realiza descartando tanto aquéllas que aparecen de manera ocasional como aquéllas que aparecen con una frecuencia muy alta. Una palabra que aparece muchas veces en distintos documentos no define ninguna categoría concreta.

Dependiendo de cuáles sean los términos resultantes tras la reducción de dimensión, se pueden distinguir dos posibles esquemas: selección de términos y extracción de términos. En el primero de ellos, el conjunto de términos resultantes $|T'|$ es un subconjunto de $|T|$. Con esta premisa, una de las funciones que más comúnmente se emplean los modelos y técnicas de RI es la de la frecuencia de los términos (número de veces que aparecen). Esto implica la necesidad de normalizar dichos términos, de manera que los recuentos de las frecuencias puedan efectuarse de manera adecuada. Dejando de lado la cuestión de las palabras vacías, hay que tener en cuenta las palabras derivadas del mismo lema, a las que cabe atribuir un contenido semántico muy próximo. Las posibles variaciones de los derivados, junto con formas flexionadas, alteraciones en género y número, etc. hacen aconsejable un agrupamiento de tales variantes bajo un único término. Lo contrario produce una dispersión en el cálculo de frecuencias de tales términos, así como la dificultad de comparar documentos [Figuerola y otros, 2004]. Esta operación se conoce como *stemming* o lematización y se emplea para reducir las dimensiones del vector que representa a un documento.

Otras funciones que se pueden utilizar en la reducción de dimensiones mediante selección de términos suelen ser x^2, coeficiente NGL, ganancia de información, información mutua y coeficiente GSS.

El segundo de los esquemas para realizar la reducción de las dimensiones del vector que representa un determinado documento es la extracción de términos. En este caso el conjunto $|T'|$ está formado por nuevos términos y no es un subconjunto de $|T|$. Debido a los problemas por polisemia, homonimia y sinonimia, los términos del vector original pueden no ser adecuados para su representación. Cualquier método de extracción de términos debe especificar la forma de extraer los nuevos términos a partir de los antiguos y la forma de convertir la representación original en nuevas representaciones basadas en los nuevos términos.

Existen dos métodos: agrupamiento de términos (*term clustering*) e indexación semántica latente [Sebastiani, 2002].

En el agrupamiento de términos se forman grupos o *clusters* con aquellos términos que presentan un elevado grado de paridad semántica, de tal forma que los *clusters*, sus centroides o cualquier posible representación de ellos se utilicen como dimensiones. Lo que se pretende es encontrar grupos de palabras que presenten relaciones semánticas en términos de su coocurrencia o coausencia en los documentos de entrenamiento. Se trata de un agrupamiento no supervisado, puesto que el agrupamiento no se ve afectado por las categorías de los documentos.

Las tareas que emplean el procesamiento del lenguaje natural, tales como la extracción de información, la búsqueda de respuestas, el resumen automático o la traducción automática, buscan resolver correctamente la variabilidad semántica. Para ello existe un área denominada Resolución de la Implicación Textual (RTE, *Recognising Textual Entailment*) [Dagan, Glickman, Magnini, 2005], que se ocupa de detectar si dos fragmentos de texto con diferente estructura tienen el mismo contenido semántico. Cada par de fragmentos se compone de una parte denominada Texto (T) y otra parte denominada Hipótesis (H). El objetivo es demostrar si el contenido del texto proporciona la misma información que el contenido de la hipótesis. Por ejemplo, $T("He\ died\ of\ blood\ loss")$ y $H("He\ died\ bleeding")$ tienen el mismo significado pero diferente estructura. Por lo tanto, se puede decir que la semántica de una sentencia se puede inferir de la semántica de la otra [Vázquez, Kozareva, Montoyo, 2006].

Por otro lado, la Indexación Semántica Latente (LSI, *Latent Semantic Indexing*), es un modelo alternativo que maneja la búsqueda de información mediante la indexación de términos, ubicándolos en un contexto semántico común [Novoa & Ballen, 2007]. La variabilidad semántica es una característica propia del lenguaje, que permite expresar de diferentes formas un mismo pensamiento. Por ello, es necesario identificar correctamente aquellas frases o fragmentos que, aunque construidos con estructuras diferentes, tienen el mismo contenido semántico [Vázquez, Kozareva, Montoyo, 2006]. Trata de resolver, por lo tanto, la implicación textual, es un modelo computacional que hace uso de la propiedad del lenguaje natural por la que palabras de igual campo semántico suelen aparecer en el mismo contexto. A partir del corpus de documentos, este modelo establece relaciones entre palabras utilizando un espacio semántico vectorial donde todos los términos son representados con una matriz.

Para obtener información útil, los términos deben estar distribuidos en documentos, párrafos o frases. Esta distribución determinará cuál es la coocurrencia entre diferentes términos y la probabilidad de utilizar otros términos en el mismo contexto. Una vez obtenida la matriz, se utiliza el Teorema de Descomposición en Valores Singulares (SVD, *Singular Value Decomposition*):

> *Sea* $A \in \mathcal{L}(R^n, R^m)$; *entonces la matriz cuadrada de dimensión* n $A^T A$ *es simétrica y semidefinida positiva; por tanto, sus valores propios* $\lambda_1, \dots, \lambda_n$ *son no negativos y existen los valores* $\sigma_i = +\sqrt{\lambda_i}$, *que se llaman valores singulares de la matriz.*
>
> *Existen además dos matrices ortogonales* U *y* V *de dimensión* m *y* n, *tales que* $U^T AV = U^{-1} AV = \Sigma$ *donde la matriz* $\Sigma \in \mathcal{L}(R^n, R^m)$ *es de la forma* $\Sigma = \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix}$ *con* $D = diag(\sigma_1, \dots, \sigma_r)$ *con* $\sigma_1 \geq \cdots \geq \sigma_r > 0$ *valores singulares de la matriz* A *no nulos y, por tanto,* r *es el rango de* A

Este teorema utiliza un algoritmo recursivo para descomponer la matriz inicial en tres nuevas matrices que contienen vectores y valores singulares [Vázquez, Kozareva, & Montoyo, 2006].

Estas matrices disminuyen el número de datos originales creando factores linealmente independientes. Una gran parte de estos factores son muy pequeños y pueden ser ignorados de forma que se obtiene un modelo aproximado reduciendo el número de factores. El resultado final es un modelo reducido de la matriz inicial que puede ser utilizado para establecer relaciones de similitud entre palabras.

2.5.1.3 ASIGNACIÓN DE PESOS

El modelo de espacio vectorial dice que un documento es considerado como un vector $d_j = \{w_{1j}, \dots, w_{|T|j}\}$, donde w_{kj} es un valor numérico que expresa la importancia de la palabra k en el documento j.

- **Representación binaria**

  La representación binaria pretende indicar si en un documento aparece (1) o no (0) una cierta palabra. La siguiente tabla resume para cada documento d_x si aparece o no el término w_n.

Documento	w_1	w_2	w_3
d_1	1	0	1
d_2	1	0	0
d_3	0	1	1
d_4	0	0	1
d_5	0	1	0

Tabla 5 - Asignación de pesos. Ejemplo de Representación Binaria

- **Frecuencia de palabra (TF, *Term Frequency*)**

  A cada palabra se le asigna una importancia proporcional al número de veces que aparece en el documento, viene dado por esta expresión:

$$w_{kj} = tf(t_k, d_j)$$

Ecuación 27 - Frecuencia de una palabra (TF)

Donde $tf(t, d)$ es la frecuencia de la palabra t en el documento d

Documento	w_1	w_2	w_3
d_1	1	2	2
d_2	4	0	3
d_3	0	5	7
d_4	7	1	2
d_5	10	1	0

Tabla 6 - Asignación de pesos. Ejemplo empleando frecuencia de palabra

En este caso, el ejemplo anterior, se indica el número de apariciones de una palabra w_n en el documento d_x.

- Frecuencia inversa del documento (IDF, *Inverse Document Frequency*)

  Aquellos términos que aparecen en gran cantidad de documentos deben ser tratados como si su relevancia fuera mínima y por lo tanto eliminarlos del vector. La importancia de cada palabra es inversamente proporcional al número de documentos que la contienen. El factor IDF para una palabra viene dado por:

$$idf(t_k) = \log\left(\frac{N}{df(t_k)}\right)$$

Ecuación 28 - IDF de una palabra

  Donde N es el número de textos totales y $df(t)$ es el número de textos que contienen el término t.

  El objetivo de este mecanismo es asignar pesos elevados a aquellas palabras poco frecuentes en los textos y pesos bajos para las palabras más comunes.

- Frecuencia de palabra por Frecuencia inversa del documento (TF·IDF)

  Cuanto más aparezca un término en un documento, más representativo será para su contenido, y cuantos menos documentos contengan dicho término, menos discriminante será para realizar la clasificación. La importancia que tiene un término para un determinado documento se calcula únicamente en función del número de apariciones. El orden en que aparece en el documento y el papel sintáctico que juega no se tiene en consideración.

$$w_{kj} = tf(t_k, d_j) \cdot id(t_k)$$

Ecuación 29 - Frecuencia de palabra por Frecuencia inversa del documento

  Típicamente se normalizan los valores dentro del intervalo [0,1] para que tengan igual longitud. La fórmula más usada es la siguiente:

$$w_{kj} = \frac{tf(t_k, d_j) \cdot id(t_k)}{\sqrt{\sum_{s=1}^{|T|}(tf(t_s, d_j) \cdot id(t_s))}}$$

Ecuación 30 - Asignación de Pesos. Normalización del coseno.

2.5.2 ENTRENAMIENTO

El entrenamiento, en un sistema de clasificación automática, implica la obtención de una fuente fiable de documentos previamente clasificados por expertos en el tema. En distintos apartados se hace referencia a este proceso de estructuración y previa clasificación de documentos.

Básicamente, el sistema aprende cómo son las categorías para después decidir a cuál de ellas se parece más el documento a clasificar.

Uno de los problemas más señalados en esta fase es el sobreentrenamiento: es el efecto de sobreajustar un algoritmo de aprendizaje con unos ciertos datos para los que se conoce el resultado deseado. Cuando un sistema se entrena demasiado o se entrena con datos extraños, el algoritmo de aprendizaje puede quedar ajustado a unas características muy específicas de los datos de entrenamiento que no tienen relación casual con la función objetivo. Durante la fase de sobreajuste el éxito al responder las muestreas de entrenamiento sigue incrementándose mientras que su actuación con muestras nuevas va empeorando.

En el siguiente ejemplo, en el caso de la línea verde, describe a la perfección los datos de entrenamiento, pero probablemente en la clasificación de documentos nuevos, el sistema de más errores que la clasificación lineal (línea negra).

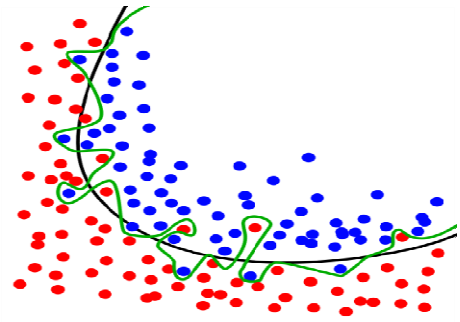


Ilustración 21 - Sobreentrenamiento

2.5.3 CLASIFICACIÓN

Una vez realizadas las fases comentadas anteriormente, la clasificación se convierte en el problema de encontrar el documento o conjunto de documentos más parecidos al dado y poder así deducir cuál es su categoría.

Un sistema de clasificación automática de vídeos, consiste, en sentido amplio, en un conjunto de algoritmos, técnicas y sistemas capaces de asignar a un vídeo una o más categorías dentro de una jerarquía dependiendo de cuál sea, por ejemplo, su afinidad temática. El escenario de aplicación de este proyecto se basa en que las clases y los grupos son determinados previamente por personas, y la labor del sistema es simplemente asignar cada documento a una de esas clases definidas a priori. Este modelo se conoce como clasificación automática supervisada, en el sentido de que requiere la supervisión o intervención humana, tanto para diseñar las clases o categorías como para entrenar el sistema.

Una vez obtenida una representación adecuada de la colección de documentos de entrenamiento, se puede aplicar un método de clasificación automática (de entre los muchos existentes) para clasificar nuevos elementos. Por tanto, la clasificación se reduce a:

Dado un documento de partida d y un modelo de clasificación entrenado E el clasificador automático deberá entonces calcular por cada una de las fases del modelo E_j cuál de ellas guarda más similitud con d.

Para ello se utilizan distintos tipos de algoritmos y técnicas explicados con anterioridad (2.4.3.3 Técnicas y algoritmos de clasificación automática de textos). En el sistema desarrollado y como se comprobará en el siguiente capítulo, se ha implementado una solución que utiliza el algoritmo K-NN modificado.

CAPÍTULO 3 - DISEÑO E IMPLEMENTACIÓN DE UN SISTEMA DE CLASIFICACIÓN DE VÍDEOS

3.1 INTRODUCCIÓN

El sistema implementado se basa en la misma arquitectura que se presentó en 2008 a VideoCLEF por Julio Villena Román y Sara Lana Serrano [Villena, Lana, 2008] pero con un módulo de transcripción automática añadido. En aquel momento, se diseñó una arquitectura y estudió un sistema basado en Lucene y Wikipedia [Wikipedia, 2009].

VideoCLEF es una tarea o apartado de CLEF que pretende investigar en sistemas y técnicas para la recuperación de información en formatos audiovisuales.

En este sistema, la base de datos del entrenamiento fue extraída del Wikipedia y la arquitectura del sistema era la mostrada en la siguiente figura.

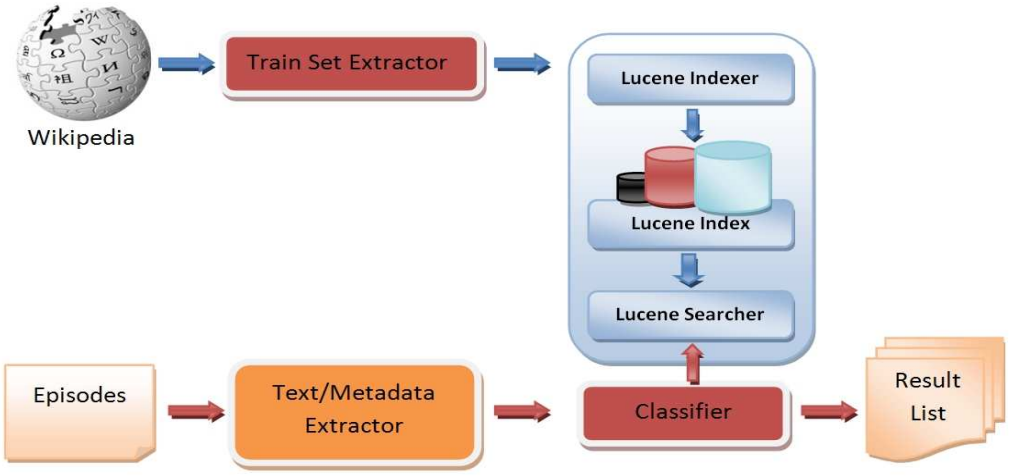


Ilustración 22 - Sistema de Clasificación (Villena y Lana)

Este sistema obtuvo la segunda posición (de 5) entre los grupos participantes y los resultados obtenidos por el sistema fueron los mostrados en la siguiente tabla (CNL,CNLEN y CNLMETA son los tres experimentos).

	Precision			Recall		
	CNL	CNLEN	CNLMETA	CNL	CNLEN	CNLMETA
Archaeology	0.25	0.25	0.40	0.14	0.14	0.29
Architecture	0.00	0.00	0.00	1.00	1.00	1.00
Chemistry	0.00	0.00	0.00	1.00	1.00	1.00
Dance	0.00	0.00	0.13	0.00	0.00	0.6
Film	1.00	0.25	1.00	0.00	0.33	0.00
History	0.25	0.26	0.38	0.30	0.50	0.60
Music	**0.64**	**0.65**	**0.65**	**0.95**	**1.00**	**1.00**
Paintings	1.00	0.00	1.00	0.00	0.00	0.00
Scientific Research	0.29	0.21	0.21	1.00	1.00	1.00
Visual Arts	1.00	0.20	0.14	0.00	0.40	0.20
ALL (microaveraged)	**0.43**	0.29	0.37	0.51	0.61	**0.65**
ALL (macroaveraged)	**0.44**	0.18	0.39	0.44	0.54	**0.58**

Tabla 7 - Resultados del Sistema de Clasificación VideoCLEF' 08 (Villena y Lana)

Lo que se pretende con este proyecto, ya expuesto con anterioridad, es desarrollar y estudiar un sistema de clasificación automática de vídeos capaz de ser totalmente automático, es decir, que sea capaz de dado un vídeo clasificarlo, haciendo su correspondiente transcripción automática y su correspondiente clasificación.

El actual sistema con el que se ha trabajado conserva la parte de algoritmos y técnicas de clasificación basados en Lucene (K-NN modificado), pero, por el contrario, modifica la base de datos y añade un módulo de transcripción automática. Es un sistema de clasificación supervisada, no paramétrico (ambos, tanto basado en patrones como en ejemplos) y centrado en la categoría.

3.2 ARQUITECTURA

Partiendo de la base del sistema de CLEF, se ha realizado el actual proyecto. La arquitectura planteada pretende estudiar la viabilidad de un sistema de clasificación de vídeos en tiempo real y para ello es evidente que, sobre la arquitectura comentada, se necesitan ciertas modificaciones. La transcripción automática o reconocimiento del habla así como la base de datos documental es el cambio más importante y significativo.

La siguiente figura pretende ilustrar los módulos cambiados y añadidos respecto a la arquitectura de dicho sistema.

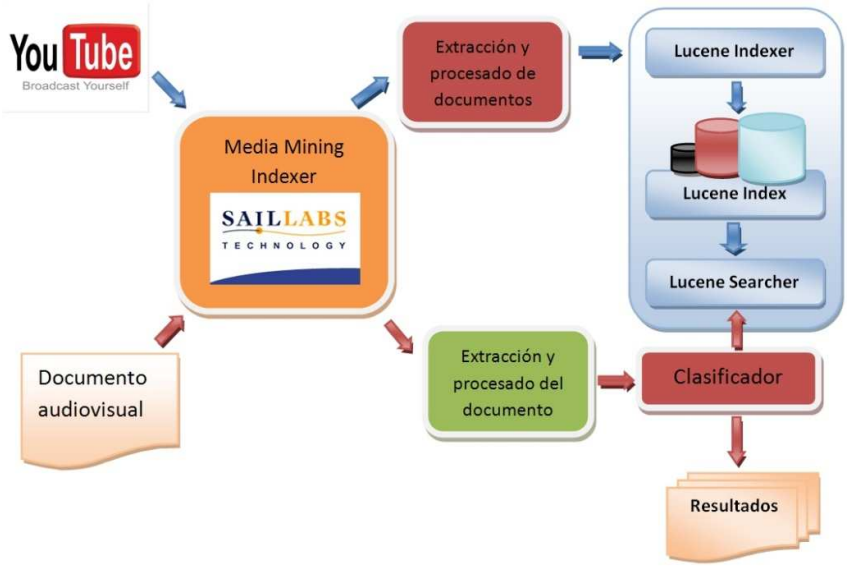


Ilustración 23 - Arquitectura del clasificador propuesto

Como se aprecia en la figura y contrastando con el sistema de Villena y Lana (Ilustración 22) la parte referente a la clasificación no cambia, es más bien el primer módulo el que modifica todo el sistema. Los vídeos son transcritos en tiempo de ejecución del clasificador y esta operación se realiza de manera automática con un sistema de reconocimiento automático del habla (Media Mining Indexer [Media Mining Indexer, 2009]) y la base de datos documental es obtenida de YouTube [YouTube, 2009].

Para comprender correctamente el funcionamiento del sistema, es conveniente describir qué hacen ciertas cajas negras del sistema.

3.2.1 MEDIA MINING INDEXER

Media Mining Indexer [Media Mining Indexer, 2009] (en adelante MMI) es un *software* de reconocimiento de voz desarrollado por Sail Labs que permite procesar voz de múltiples fuentes en varios formatos y produce texto en tiempo real, además es un sistema independiente del hablante. Por otra parte, es una herramienta de fácil uso.

MMI permite manejar variaciones en los estilos y los dominios en los que se va a realizar la grabación de voz y posee una arquitectura distribuida que permite la integración en distintas infraestructuras.

El etiquetado de tiempo de cada palabra permite identificar y acceder rápidamente a los segmentos de interés, además es capaz de realizar monitorización de medios de comunicación de una manera sencilla mediante la identificación y categorización del audio llegando incluso a la monitorización de noticias simultánea en varios idiomas, como español, inglés, francés, alemán, ruso y árabe.

En primer lugar, realiza una segmentación del audio en secciones de habla y silencio, posteriormente aplica reconocimiento del habla sobre los segmentos identificados como habla. Este proceso se realiza en tiempo real, para grandes vocabularios y datos de audio. Utiliza coeficientes MFCC y varios tipos de normalización y cancelación activa de ruido, además los modelos acústicos son independientes del hablante y su género. Los modelos del lenguaje están basados en $n - gramas$.

Las características más importantes del sistema son [Media Mining Indexer, 2009](extraído de Collada [Collada Pérez, 2009]):

- Permite una conversión precisa de habla espontanea a textos incluso con altos niveles de ruido de fondo.
- Detección del cambio del hablante analizando segmentos de audio para la recuperación de información.
- Identificación del hablante.
- Detección del tema.
- Detección de entidades que permiten encontrar información relevante, como palabras pertenecientes a categorías de tipo personas, lugares, organizaciones, etc.
- Traducción de palabras clave, permite acceso inmediato a información contenida en otros idiomas.
- Arquitectura escalable que permite tanto el uso en un único ordenador como la configuración multimáquina, proporcionando una solución de coste efectivo para estaciones de televisión de tamaño pequeño o mediano.
- La salida en formato XML permite la integración con tecnologías complementarias.
- Los idiomas disponibles son: Árabe, Inglés (Estados Unidos), Inglés (Estados Unidos/Reino Unido) Francés, Alemán, Griego, Noruego, Ruso, Español.

3.2.2 LUCENE

Lucene es una librería utilizada como herramienta para incorporar capacidades de recuperación de información a las aplicaciones. Ofrece alto rendimiento, eficiencia y es muy robusta y versátil.

Por otro lado, es una herramienta de *software* libre creada por Doug Cutting que es el creador de Nutch, primer buscador abierto y en la actualidad miembro de la Apache Software Foundation [Apache, 2009].

Lucene está catalogado como una librería de calidad industrial realizada en Java [Java, 2009] y gratuita. Las características más importantes de Lucene son [Paz, 2008]:

- Está distribuida bajo la licencia Apache, que permite su uso en aplicaciones tanto de código abierto como comerciales.

- Es multiplataforma en su versión Java o migrada a diferentes lenguajes de programación (C/ C++, .NET, Python, Ruby, PHP, etc.).
- Está en constante mantenimiento y actualización, la última versión publicada es la Lucene.
- Java 2.9.0 [Lucene, 2009].
- Incluye algoritmos de búsqueda potentes, fiables y eficientes:
    - Lenguaje de consulta muy potente (literales, frases, comodines, búsqueda por proximidad, rangos, búsqueda aproximada, etc.)
    - Búsqueda por campos de documento.
    - Búsqueda por rango de fechas.
    - Resultados ordenados por relevancia y/o cualquier campo.
    - Búsqueda multi-índice en paralelo y combinación de resultados.
    - Permite búsquedas a la vez que se está indexando y optimizando el índice.
- Ofrece un muy alto rendimiento y es fácilmente escalable[3]
    - Más de 20MB/minuto en Pentium M 1.5GHz.
    - Bajos requerimientos de memoria RAM (sólo 1MB de *heap*).
    - Indexación incremental tan rápida como por lotes.
    - El índice sólo ocupa el 20-30% del texto indexado.
    - Capaz de manejar ingentes cantidades de datos.

3.3 OBTENCIÓN DEL CORPUS

El primer paso para crear el sistema es la obtención del corpus de vídeos. Como base de datos para este proyecto se ha utilizado YouTube, tanto por su gran accesibilidad como por su gran cantidad de vídeos catalogados previamente.

YouTube ofrece diversas categorías catalogadas por usuarios externos y mantiene una fiabilidad bastante aceptable para ser una base documental abierta y libre. Permite la descarga y posee aplicaciones en varios lenguajes de programación para su integración en otros sistemas, así como desarrollo de aplicaciones. Una característica importante de los vídeos de YouTube es que poseen una ficha con información complementaria sobre el vídeo, con la descripción, la fecha, el autor y la categoría entre otros.

Los ficheros son descargados en formato vídeo, después se extrae su audio, se procesan y se convierten a texto con MMI creándose un fichero XML. Esta salida es procesada y extraída en varios tipos, ya que, como se verán más adelante, las pruebas serán realizadas de diferentes maneras. Para realizar los procesos descritos se han implementado ciertas clases y *scripts* en PHP [PHP, 2009], un esquema general, ya que no es el objetivo de este estudio, sería el siguiente:

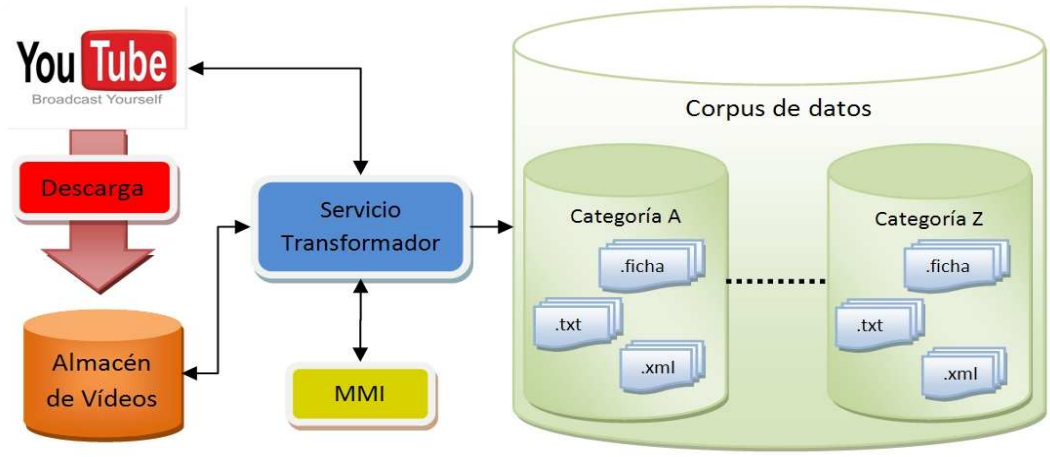


Ilustración 24 - Procesado de vídeos. Obtención del corpus

Los ficheros de vídeo se descargan en formato FLV, después se les extrae el audio con el programa MPlayer [MPlayer, 2009] en formato WAV. Posteriormente, los ficheros de audio son convertidos en 3 ficheros distintos:

- **.ficha:** Contiene la información referente al vídeo. Características tipo etiquetas, categorías, fechas, etc. Se consigue extrayendo la información de la URL de cada vídeo.
- **.xml:** Fichero devuelto por MMI. Contiene datos referentes al tiempo y secuencia de cada palabra, así como tipo de hablante y otros.
- **.txt:** Texto del documento audiovisual. Una vez obtenido el xml de cada vídeo, es procesado y se extrae su texto.

En apartados siguientes se explicará con más detalle, pero es importante destacar, llegado a este punto, que del fichero xml devuelto por MMI se extraen varios tipos de ficheros .txt para las posteriores pruebas.

Los criterios establecidos son: cien palabras más frecuentes del documento, treinta primeros segundos de cada documento y documento completo para cada una de las 13 categorías.

El corpus anteriormente comentado fue extraído en marzo de 2009. Existían otra categoría, en concreto Música, que fue eliminada del corpus porque era incompatible. La razón es que no hablaba de música si no que contenía videoclips y otros de este estilo. Lógicamente el sistema no puede incorporar esta categoría.

Realizada la operación anterior, el sistema creado obtiene la siguiente base de datos de vídeos:

Categoría	Entrenamiento	Test
Automoción	100	20
Ciencia y tecnología	100	20
Cine y animación	100	20
Comedia	100	20
Deportes	100	20
Educación	100	20
Gente y blogs	100	20
Instrucciones varias y estilo	100	20
Juegos	100	20
Mascotas y animales	100	20
Noticias y política	100	20
Ocio	100	20
Viajes y eventos	100	20

Tabla 8 - Conjunto de Entrenamiento. Vídeos por categoría

3.4 ENTRENAMIENTO

El clasificador ya tiene, entonces, la lista de documentos que requiere para crear su entrenamiento. El sistema será entrenado con cada tipo de fichero previamente dicho y será probado con los mismos, es decir, que el clasificador será probado con tres tipos distintos de corpus de entrenamiento.

Entrenamiento	Tipo de ficheros
100Palabras	Estos ficheros contienen las cien palabras más frecuentes del documento.
30Segundos	El entrenamiento se realiza con los treinta primeros segundos del vídeo
DocumentoCompleto	Corresponde al entrenamiento y pruebas realizadas con los documentos completos.

Tabla 9 - Tipos de Entrenamiento

La indexación, estructuración y preprocesado de cada documento es llevado a cabo por Lucene. Concretamente, cada documento es tratado con los siguientes pasos:

- **Extracción del texto:** Ya comentado anteriormente, dado el fichero XML de MMI el texto es extraído filtrando únicamente las palabras, en el orden correcto. Esto es repetido en cada uno de los tipos de ficheros, con las peculiaridades de cada uno, que han sido comentadas en la Tabla 9 - Tipos de Entrenamiento.
- **Eliminación de signos diacríticos y la conversión a minúsculas:** Todos los términos están normalizados mediante la eliminación de los signos diacríticos y cambiar todas las letras en minúsculas.

- **Filtrado:** Todas las palabras reconocidas como palabras vacías son eliminadas. Las palabras vacías son recogidas por Villena y Lana de sus propios conocimientos y recursos.
- **Lematización**: Se realiza una lematización por cada término del documento para ayudar en la recuperación de información. Se ha usa el *standar stemmer* de Porter [Porter, 2008].

Por último, decir que el sistema ha implementado dos tipos de entrenamiento. Es un sistema basado en parámetros estadísticos, implementa una clasificación no paramétrica (ver 2.4.3.2 Tipos de clasificadores automáticos) y se ha realizado tanto basado en patrones como en ejemplos. Es decir, por cada tipo de entrenamiento (Tabla 9 - Tipos de Entrenamiento) se ha realizado la indexación tanto basándose en ejemplos (n documentos por categoría) como en patrones (1 documento por categoría).

3.5 CLASIFICACIÓN

El sistema, como se ha dicho repetidas veces, está basado en Lucene que implementa el algoritmo de K-NN modificado. Para encontrar una clase o categoría a un documento, el sistema realiza las siguientes operaciones:

- **Transcripción automática del documento:** Tras extraer el audio del documento se realiza su transcripción a texto. En el caso de las pruebas también se extraerá información de YouTube para posteriormente comprobar si se clasificó bien o no.
- **Preprocesado del documento**: Una vez que se tiene el xml dado por MMI en la transcripción automática, se extrae el texto. Como en el entrenamiento, se eliminan signos, se filtra y se lematiza. Hay que recordar que dependiendo del tipo de prueba que se esté realizando, además, el documento es filtrado por otro proceso (ver Tabla 9 - Tipos de Entrenamiento).
- **Búsqueda de la categoría:** Se realiza entonces la consulta a Lucene para que devuelva la lista de documentos relevantes. Si la clasificación o la prueba está basada en ejemplos, previamente se normalizan sus puntuaciones y se tiene en cuenta tanto el número de documentos de una categoría como la posición y puntuación que da Lucene. Si se clasifica basándose en patrones, Lucene devuelve únicamente un documento por cada categoría que es similar o se parece al documento.

3.6 EVALUACIÓN

La evaluación del sistema es la parte más importante de este proyecto y es por eso por lo que se le ha dedicado un capítulo entero, Capítulo 4.

CAPÍTULO 4 - EVALUACIÓN

4.1 INTRODUCCIÓN

En el análisis de resultados se pretende evaluar el sistema. Para ello se han realizado diferentes pruebas independientes, dependiendo del tipo de entrenamiento y de los índices (tipos de documentos) como se detalla en la Tabla 10 - Tipos de entrenamiento en la evaluación. Un detalle a tener en cuenta, es que, no se ha evaluado la calidad de la transcripción automática ya que es una caja negra para este sistema. Se da por hecho, entonces, que la clasificación es realizada correctamente.

Cada prueba será estudiada por separado haciendo un análisis objetivo y mostrando la evolución de los resultados al ir cambiando los distintos parámetros del sistema.

A modo de recordatorio es importante recordar los siguientes conceptos y alias utilizados:

- **Tipos de clasificación:**
  - Clasificación basada en patrones: Descripción de cada clase o categoría en términos de un patrón y comparación del texto con dichos patrones.
  - Clasificación basada en ejemplos: Se basa en la similitud con un conjunto de ejemplos.
- **Tipos de documentos:**
  - Documentos completos: Transcripción completa de los documentos.
  - Treinta primeros segundos: Transcripción de los documentos acotada a los 30 primeros segundos de cada vídeo.
  - Cien palabras más repetidas: Se refiere a las cien palabras que más aparecen en el documento.
- **Palabras:** Cuando en las pruebas se habla de palabras, se hace referencia al número de palabras o términos escogidos por el clasificador como palabras importantes o que mejor representan a un vídeo.
- **Vídeos o vecinos:** Se refiere al número de vecinos más cercanos que devuelve el sistema. Recordar que el sistema implementa una versión mejorada del algoritmo K-NN (K vecinos más próximos).
- **M y m:** Se refiere respectivamente, a macro-averaging y micro-averaging, medidas de evaluación explicadas anteriormente (2.4.4.3 Evaluación de los sistemas de clasificación automática de).

La evaluación del sistema es la parte más importante de este proyecto y es por eso por lo que se le ha dedicado un capítulo entero. A continuación se va hacer una introducción y un

esquema general de todas las pruebas realizadas. La siguiente tabla resume todos los datos y tipos de entrenamiento realizados:

Entrenamiento	Tipo de clasificación	Tipo de documento
	Basado en Patrones	*100Palabras*
		30Segundos
		DocumentoCompleto
	Basado en Ejemplos	*100Palabras*
		30Segundos
		DocumentoCompleto

Tabla 10 - Tipos de entrenamiento en la evaluación

Es decir, que el sistema tiene 6 tipos de índices, y por lo tanto, se han realizado 6 tipos de pruebas, una por cada tipo de entrenamiento realizado.

Por otra parte, con cada entrenamiento se han probado de diferentes maneras, ya que, por ejemplo, en el caso de la clasificación basada en ejemplos el número de vídeos que se quiere que devuelva Lucene al clasificar un documento es relevante para el resultado final. También es importante el número de palabras que se considerarán importantes en un documento. Ambos son parámetros de entrada de Lucene.

Por lo tanto, y expandiendo la tabla anterior, las pruebas realizadas serían las siguientes.

Tipo de Clasificación	Tipo de documento	Nº de Palabras	Nº de Documentos	Nº Documentos Test
Basado en Patrones	*100Palabras*	Si	No	260
	30Segundos	Si	No	260
	DocumentoCompleto	Si	No	260
Basado en Ejemplos	*100Palabras*	Si	Si	260
	30Segundos	Si	Si	260
	DocumentoCompleto	Si	Si	260

Tabla 11 - Pruebas realizadas en la evaluación

En la tabla anterior se habla de "número de palabras" y de "número de documentos". Cuando se hace referencia a esto, quiere decir que se han realizado pruebas cambiando dichos parámetros. El número de palabras ronda los valores de 10 a 200 en intervalos de 10 en 10 $\{10,20,30,\dots,200\}$ mientras que el número de documentos se mueve en el intervalo $\{10, 15, \dots 50\}$, entendiendo documentos como vídeos.

Para cada prueba realizada se han calculado tanto la precisión como la cobertura, de manera global y para cada categoría.

4.2 LOS RESULTADOS

4.2.1 CLASIFICACIÓN BASADA EN PATRONES

Las clases o categorías se representan con un vector de términos y la aplicación pretende averiguar a cuál de estos vectores se parece más el documento de test. El conjunto de pruebas está formado por 260 vídeos, repartidos en 20 vídeos de cada una de las 13 categorías y se repiten para los distintos tipos de documentos.

Los resultados obtenidos por estas pruebas están detallados en la gráfica correspondiente al Anexo 2 y en la tabla del Anexo 1.

La siguiente tabla muestra un resumen de los mejores resultados obtenidos en cada este tipo de clasificación (m micro-averaging, M macro-averaging, y en el resto de gráficas será igual). Como mejor clasificación se ha tomado aquella que más precisión tiene a la primera (en P1).

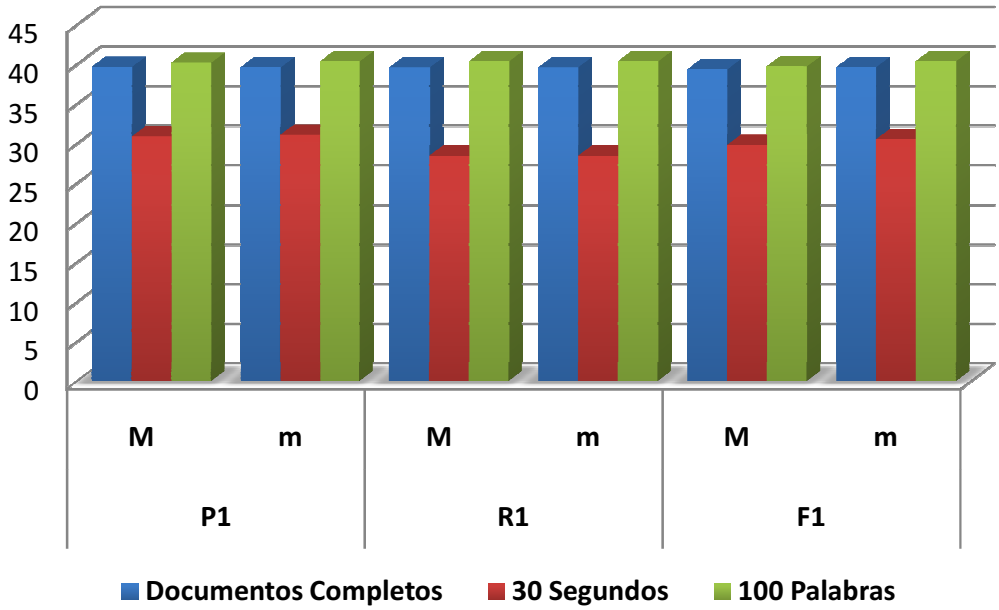
Tipo de documento	Palabras	P1		R1		F1	
		M	m	M	m	M	m
Documentos	50	39,68	39,62	39,615	39,615	39,36	39,62
Completos	60	38,47	38,46	38,46	38,46	38,19	38,46
30 Segundos	50	30,88	31,09	28,46	28,46	29,81	30,53
	60	30,88	31,09	28,46	28,46	29,81	30,53
100	50	40,34	40	40	40	39,86	40
Palabras	**60**	**40,21**	**40,38**	**40,38**	**40,38**	**39,73**	**40,38**

Tabla 12 - Mejores resultados - Clasificación basada en patrones (precisión en 1)

Tanto las pruebas realizadas con documentos completos como las realizadas con las cien palabras más repetidas destacan por su mayor precisión, cobertura y medida-F. Los resultados obtenidos demuestran que el tamaño de los documentos influye en la efectividad del clasificador.

Documentos pequeños, como los que proporciona el conjunto de test formado por los 30 primeros segundos de cada documento dan peores resultados que aquellos que, o bien tienen todas las palabras o bien tienen las palabras que más aparecen en el documento.

La siguiente gráfica representa una comparativa de los mejores resultados.



Gráfica 1 - Diagrama de bloques - Clasificación basada en patrones

En los siguientes apartados se detallan las pruebas realizadas con los distintos índices o tipos de documentos para la clasificación basada en patrones.

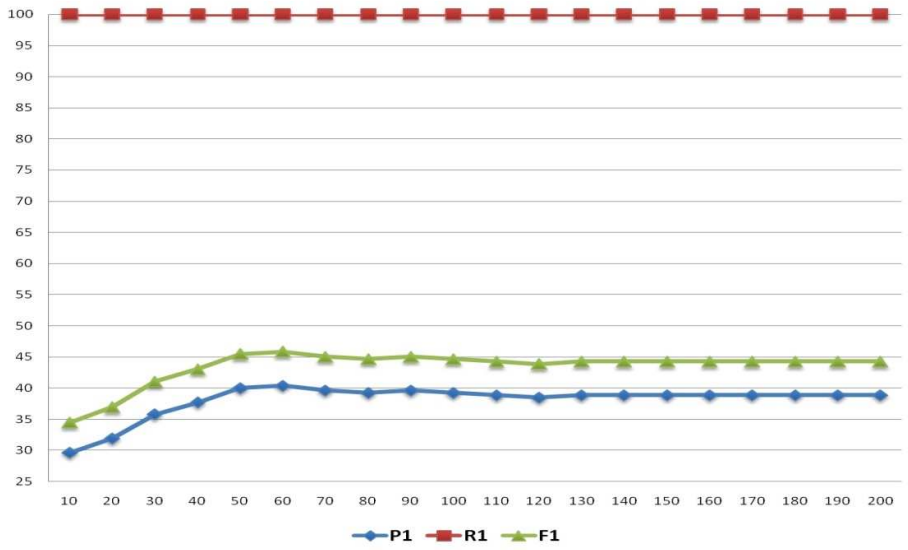
4.2.1.1 RESULTADOS DE LA CLASIFICACIÓN CON LAS CIEN PALABRAS MÁS IMPORTANTES

Los resultados obtenidos para este tipo de entrenamiento están los detallados en la tabla del Anexo 1. De manera que resuma las anteriores tablas, se han escogido los diez mejores resultados para poder hacer una comparativa y explicar a la vez que analizar los resultados obtenidos.

La siguiente tabla muestra la precisión, la cobertura y la medida-F para los resultados obtenidos a la primera, es decir, precisión en uno (P1), cobertura en uno (R1) y medida-F en uno (F1), con respecto al número de palabras.

Palabras	P1	R1	F1
50	40	100	45,45
60	**40,38**	**100**	**45,85**
70	39,61	100	45,06
80	39,23	100	44,66
90	39,61	100	45,06
100	39,23	100	44,66
170	38,84	100	44,26
180	38,84	100	44,26
190	38,84	100	44,26
200	38,84	100	44,26

Tabla 13 - Resultados de la Clasificación basada en patrones - Cien palabras más repetidas



Gráfica 2 - Resultados de la Clasificación basada en patrones - Cien palabras más repetidas

Se puede observar que el sistema tiene una cobertura perfecta, es decir, clasifica todos los vídeos del conjunto de test, según la fórmula definida anteriormente. Por otra parte, se observa que los mejores valores se concentran alrededor de los valores 50 y 60 (palabras), al tener una precisión tan baja, la medida de ponderación entre la precisión y la cobertura (medida-F) se aproxima a la precisión, es decir, es relativamente baja.

Palabras	P1		R1		F1	
	M	m	M	m	M	m
10	28,08	29,62	29,62	29,62	28,17	29,62
20	30,55	31,92	31,92	31,92	30,53	31,92
30	34,29	35,77	35,77	35,77	34,16	35,77
40	36,57	37,69	37,69	37,69	36,41	37,69
50	**40,34**	**40**	**40**	**40**	**39,86**	**40**
60	**40,21**	**40,38**	**40,38**	**40,38**	**39,73**	**40,38**
70	39,79	39,62	39,62	39,62	39,22	39,62
80	39,37	39,23	39,23	39,23	38,76	39,23
90	39,89	39,62	39,62	39,62	39,36	39,62
100	39,65	39,23	39,23	39,23	38,99	39,23
110	39,14	38,85	38,85	38,85	38,48	38,85
120	38,71	38,46	38,46	38,46	38,04	38,46
130	39,32	38,85	38,85	38,85	38,64	38,85
140	39,32	38,85	38,85	38,85	38,64	38,85
150	39,17	38,85	38,85	38,85	38,52	38,85
160	39,17	38,85	38,85	38,85	38,52	38,85
170	39,14	38,85	38,85	38,85	38,48	38,85
180	39,14	38,85	38,85	38,85	38,48	38,85
190	39,14	38,85	38,85	38,85	38,48	38,85
200	39,22	38,85	38,85	38,85	38,54	38,85

Tabla 14 - Macro-Averaging y Micro-Averaging - Clasificación basada en patrones - Cien palabras más repetidas

La anterior tabla muestra las medidas de micro-averaging y macro-averaging de manera más concreta y centrándose en los resultados obtenidos para cada categoría.

Se aprecia que la mejor puntuación coincide con la anteriormente citada, es decir, que aparentemente el mejor comportamiento del clasificador es aquella que valora las 60 palabras más importantes (aunque muy similar a la configuración con 50 palabras). Al igual que anteriormente, la precisión aumenta al aumentar el número de palabras, hasta llegar a 60 donde disminuye.

Es importante fijarse en la cobertura del sistema claramente distinta a los valores anteriores, la razón de esta variación es que la cobertura es una medida que, en aspectos globales, evalúa el grado en que el sistema es capaz de clasificar un documento en cualquier categoría, mientras que, centrándose en las categorías, la cobertura mide la capacidad del sistema para clasificar un documento en una categoría concreta. Por lo tanto, el sistema tendrá una cobertura alta, si se mide en términos globales, ya que el sistema siempre que clasifique un vídeo hará subir su cobertura, mientras que, en términos de categorías concretas, el sistema tiene una cobertura baja ya que, solo subirá la cobertura para una categoría si ese documento ha sido clasificado bien.

Centrándose, entonces, en esta configuración del sistema, las medidas asociadas a las categorías son las mostradas en la tabla.

Categorías	FN	FP	TP	P1	R1	F1
Automoción	17	13	3	18,75	15	17,86
Ciencia y tecnología	4	10	16	61,54	80	64,52
Cine y animación	16	17	4	19,05	20	19,23
Comedia	14	19	6	24	30	25
Deportes	15	18	5	21,74	25	22,32
Educación	9	7	11	61,11	55	59,78
Gente y blogs	17	9	3	25	15	22,06
Instrucc. varias y estilo	6	3	14	82,35	70	79,55
Juegos	**2**	**3**	**18**	**85,71**	**90**	**86,54**
Mascotas y animales	16	6	4	40	20	33,33
Noticias y política	8	20	12	37,5	60	40,54
Ocio	15	16	5	23,81	25	24,04
Viajes y eventos	16	14	4	22,22	20	21,74
Total	155	155	105	-	-	-
			Macro-Averaging	**40,21**	**40,3846**	**39,73**
			Micro-Averaging	**40,38**	**40,3846**	**40,38**

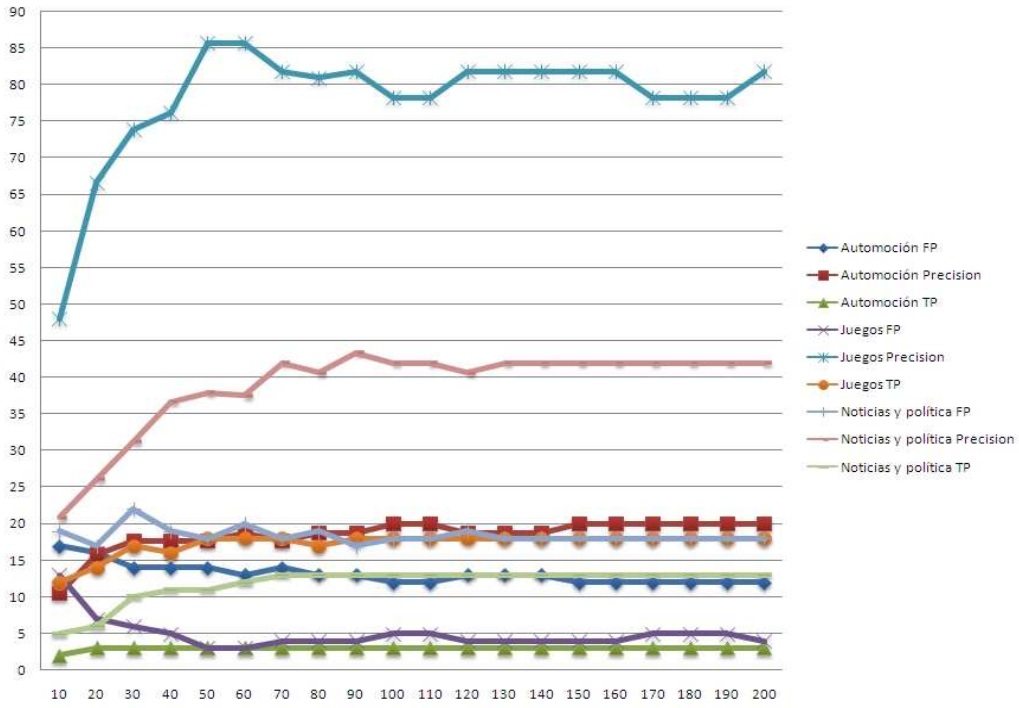
Tabla 15 - Características de la evaluación por categorías - Clasificación basada en patrones - Cien palabras más repetidas

Como se puede apreciar, ciertas categorías (Automoción, Deportes y Cine entre otras) poseen una precisión baja debido al gran número de FP. Se aprecia también, que parece haber una relación bastante clara entre el número de FN y el número de FP, apareciendo en bastantes ocasiones como números muy parecidos, de ahí el parecido entre la cobertura y la precisión. Se podría decir entonces que el sistema, ya que para el caso contrario parece también corroborarse,

posee una relación entre la capacidad para clasificar un documento y cantidad que clasifica bien, como bien demuestran las correspondientes macro y micro averaging al tener unas medidas similares.

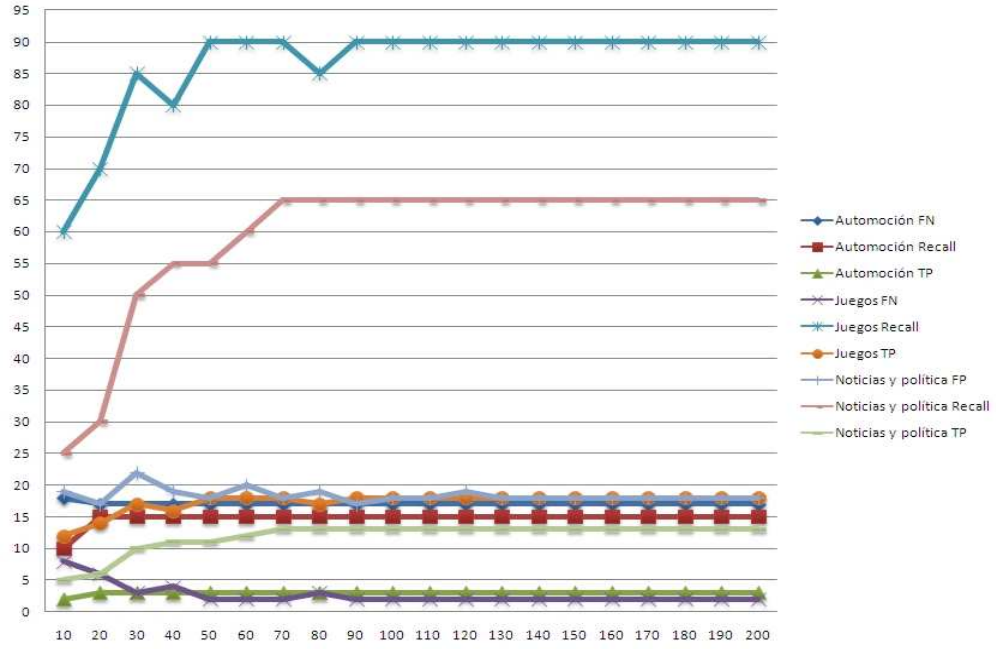
Es importante observar la evolución de los datos para cada categoría a medida que cambian las palabras. Observar la P1, la R1 y F1, viendo cómo cambia a la vez la configuración del sistema. Comparar el número de TP frente a los FP al igual que frente a los FN haciendo, a su vez, la comparativa con la precisión y la cobertura correspondiente (macro y micro averaging).

Para generar estas gráficas se han escogido tres categorías, la que tiene la precisión más alta (Juegos), la que tiene la precisión más baja (Automoción) y otra intermedia (Noticias).



Gráfica 3 - Clasificación basada en patrones - Precisión de las categorías (P1)

Mientras que el número de TP de la categoría Automoción no supera 5, se puede observar como el número de FP de Juegos casi lo iguala, se observa pues, la estrecha relación entre la precisión y estas dos medidas.



Gráfica 4 - Clasificación basada en patrones - Cobertura en las categorías (R1)

Lo siguiente que se va a analizar, es la manera en que el sistema confunde las categorías entre sí. Para ello es necesario estudiar la matriz de confusión para la configuración comentada y observada hasta ahora.

Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**3**	0	5	0	5	1	1	0	1	1	1	2	0
Cienc.	0	**16**	0	0	1	0	0	0	1	0	2	0	0
Cine	4	0	**4**	3	5	1	0	0	0	0	0	1	2
Come.	0	0	0	**6**	1	0	2	0	0	0	5	6	0
Dep	0	1	0	2	**5**	1	2	1	0	0	2	5	1
Educ.	2	0	2	2	0	**11**	0	0	0	0	3	0	0
Gent	4	0	2	4	1	1	**3**	1	0	0	1	1	2
Inst.	0	0	1	1	0	0	1	**14**	0	0	0	1	2
Jue.	0	0	0	0	0	0	1	0	**18**	0	1	0	0
Masc	0	6	1	0	2	1	0	0	1	**4**	4	0	1
Not.	1	1	0	2	1	0	2	0	0	0	**12**	0	1
Ocio	1	1	1	3	2	1	0	1	0	0	0	**5**	5
Viaj.	1	1	5	2	0	1	0	0	0	5	1	0	**4**

Tabla 16 - Matriz de Confusión - Clasificación basada en patrones - Cien palabras más repetidas

La diagonal principal (en negrita) de la matriz coincide con el número de TP dados en la tabla comentada anteriormente (Tabla 15), en horizontal se aprecian resultados reales, es decir, en este caso, de 20 documentos que hay de cada categoría, el sistema (fijándose en la primera fila) predijo que 3 eran de automoción y el resto lo confundió (un total de 17), mientras que, 13 documentos de los 260 dijo que eran de automoción siendo esto falso.

Se observa que puede distinguir casi perfectamente entre vídeos de Juegos o Ciencia mientras que, por ejemplo, distingue y confunde entre los vídeos de Ocio y Comedia. Se aprecia claramente que Viajes y eventos, quizás por su propia naturaleza como clase (ya que puede hablar de casi cualquier otra categoría), la confunde con casi todas las demás categorías.

4.2.1.2 RESULTADOS DE LA CLASIFICACIÓN CON LOS TREINTA PRIMEROS SEGUNDOS

Siguiendo la misma estructura que en el apartado anterior, se van a analizar las pruebas realizadas con este tipo de documentos. Como ya se comentó, ahora se pretende comprobar si es viable clasificar un vídeo obteniendo simplemente los 30 primeros segundos del documento.

El comportamiento del sistema con esta clase es extraño, en el sentido de que, su precisión o cobertura no varían de manera global (Anexo 1). De las 20 pruebas realizadas comentadas en el capítulo anterior (de 10 a 200 palabras variando de 10 en 10), 19 obtienen el mismo resultado, frente a una que es menor. La siguiente tabla resume este comportamiento:

Palabras	P1	R1	F 1
10	28,99	75,82	33,08
Resto	31,09	77,08	35,31

Tabla 17 - Resultados de la Clasificación basada en patrones - Treinta primeros segundos

Cabe destacar, en esta prueba, además del comportamiento extraño, la baja cobertura del sistema en términos globales (Tabla 12), lo que indica que desde el principio existen documentos que no son clasificados en ninguna categoría. Este comportamiento también se ve reflejado cuando se analizan las categorías una a una, lo que permite resumir esta prueba en la siguiente tabla:

Categorías	FN	FP	TP	P1	R1	F1
Automoción	13	8	7	46,67	35	43,75
Ciencia y tecnología	7	11	13	54,17	65	56,03
Cine y animación	18	6	2	25	10	19,23
Comedia	18	13	2	13,33	10	12,5
Deportes	19	18	1	5,263	5	5,208
Educación	13	24	7	22,58	35	24,31
Gente y blogs	19	7	1	12,5	5	9,615
Instrucc. varias y estilo	**9**	**4**	**11**	**73,33**	**55**	**68,75**
Juegos	9	10	11	52,38	55	52,88
Mascotas y animales	17	12	3	20	15	18,75
Noticias y política	11	13	9	40,91	45	41,67
Ocio	16	12	4	25	20	23,81
Viajes y eventos	17	26	3	10,34	15	11,03
Total	186	164	74	-	-	-
			Macro-Averaging	**30,88**	**28,46**	**29,81**
			Micro-Averaging	**31,09**	**28,46**	**30,53**

Tabla 18 - Características de la evaluación por categorías - Clasificación basada en patrones - Treinta primeros segundos

También en este caso se aprecia la estrecha relación entre la cobertura y la precisión. Es importante analizar el total de falsos negativos, es decir, aquellos que no fueron clasificados en su correspondiente categoría, que es, mayor que el número de fallos. Es decir, que existe un total de 22 (vídeos totales – TP – FP) documentos que no se han clasificado en ninguna categoría. De ahí se explica la baja cobertura comentada con anterioridad. Eligiendo una configuración cualquiera, ya que son iguales, como queda claro en la Tabla 17, de las que están comprendidas entre 20 y 200 palabras, la matriz de confusión (puesto que son iguales) queda de la siguiente manera.

Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**7**	0	1	1	3	2	1	0	0	1	1	1	1
Cienc.	1	**13**	0	0	0	2	0	1	0	0	0	0	2
Cine	0	1	**2**	2	1	1	0	0	2	1	3	0	1
Come.	0	0	1	**2**	3	3	1	0	1	3	0	2	3
Dep	0	1	0	0	**1**	2	1	1	2	2	0	3	0
Educ.	1	0	1	1	2	**7**	1	0	1	0	1	0	5
Gent	1	3	0	1	0	1	**1**	0	0	2	3	4	3
Inst.	0	0	0	3	0	2	1	**11**	0	2	0	0	1
Jue.	3	1	0	1	1	2	0	0	**11**	0	1	0	0
Masc	2	1	0	1	1	2	1	0	0	**3**	2	0	4
Not.	0	2	1	0	1	2	0	1	1	0	**9**	1	1
Ocio	0	0	1	1	4	2	1	0	1	1	0	**4**	5
Viaj.	0	2	1	2	2	3	0	1	2	0	2	1	**3**

Tabla 19 - Matriz de Confusión - Clasificación basada en patrones - Treinta primeros segundos

La matriz es bastante compleja, además de observarse que en la gran mayoría de las categorías no se clasifican todos los documentos, se observa el bajo índice de acierto por parte del sistema.

También parece complicado definir entre qué categorías se confunde el sistema, ya que, observando los datos, la dispersión de los mismos, es bastante uniforme. Es el caso de la categoría Viajes, donde la mayoría de los documentos han sido clasificados indistintamente entre el resto de categorías.

4.2.1.3 RESULTADOS DE LA CLASIFICACIÓN CON EL DOCUMENTO COMPLETO

Por último, se van a analizar las pruebas referentes a los documentos completos, este sistema representa al que más se asemeja con un comportamiento natural, pero no implica que deba ser mejor.

Las medidas que caracterizan a este sistema son las mostradas en la tabla.

Palabras	P1	R1	F1
10	31,54	100	36,54
20	32,69	100	37,78
30	36,92	100	42,25
40	37,69	100	43,06
50	**39,62**	**100**	**45,06**
60	38,46	100	43,86
70	37,69	100	43,06
80	38,08	100	43,46
90	38,08	100	43,46
100	37,69	100	43,06
110	37,31	100	42,66
120	36,92	100	42,25
130	36,92	100	42,25
140	36,92	100	42,25
150	36,92	100	42,25
160	36,92	100	42,25
170	37,31	100	42,66
180	37,31	100	42,66
190	37,31	100	42,66
200	37,31	100	42,66

Tabla 20 - Resultados de la Clasificación basada en patrones - Documentos Completos

La mayoría de los resultados muestran valores distintos, no obstante, casi todos rondan un rango muy similar. Una cobertura tan alta sólo puede indicar que todos los documentos son clasificados en alguna categoría. Cabe destacar que estas medidas son a partir del resultado en uno.

En todas las pruebas, para calcular la medida-f se está utilizando un $\beta = 0.5$ dando la misma importancia tanto a la cobertura como a la precisión.

Es importante darse cuenta que es evidente que la cobertura no puede tener una importancia superior a la precisión ya que, sus valores son poco representativos en estas tablas, es lógico que el clasificador tienda a clasificar el máximo número de documentos. Mas debido a la poca precisión que se tiene y a que es la medida que hasta ahora está caracterizando al sistema, tampoco tiene sentido hacer que la medida-F sea favorecida por la precisión.

Para el caso remarcado, es decir, con 50 palabras, las medidas características referentes a las categorías son bastantes similares, exceptuando claro, el caso de la cobertura.

Categorías	FN	FP	TP	Precisión	Recall	Medida-F
Automoción	17	12	3	20	15	18,75
Ciencia y tecnología	4	9	16	64	80	66,67
Cine y animación	15	16	5	23,81	25	24,04
Comedia	14	21	6	22,22	30	23,44
Deportes	15	12	5	29,41	25	28,41
Educación	9	7	11	61,11	55	59,78
Gente y blogs	17	10	3	23,08	15	20,83
Instrucc. varias y estilo	**7**	**2**	**13**	**86,67**	**65**	**81,25**
Juegos	2	6	18	75	90	77,59
Mascotas y animales	13	10	7	41,18	35	39,77
Noticias y política	12	18	8	30,77	40	32,26
Ocio	17	19	3	13,64	15	13,89
Viajes y eventos	15	15	5	25	25	25
Total	157	157	103	-	-	-
Macro-Averaging				**39,68**	**39,62**	**39,36**
Micro-Averaging				**39,62**	**39,62**	**39,62**

Tabla 21 - Características de la evaluación por categorías - Clasificación basada en patrones - Documentos Completos

Aquí se puede observar el porqué de la cobertura 100 en la tabla anterior ya que todos los documentos, como indican los totales, son clasificados en al menos una categoría ($FN + TP = Documentos\ totales$). No obstante, eso únicamente se puede aplicar a la prueba en general, como se ha comentado varias veces, sin embargo, en cuanto a categorías, el resultado es otro, ya que existen muchos documentos no clasificados en la categoría a la que corresponden.

Una vez más la categoría de Instrucciones es la que mejor resultados da en términos de P1, es preciso analizar bien la matriz de confusión para esta prueba para poder así ver el comportamiento del sistema al completo.

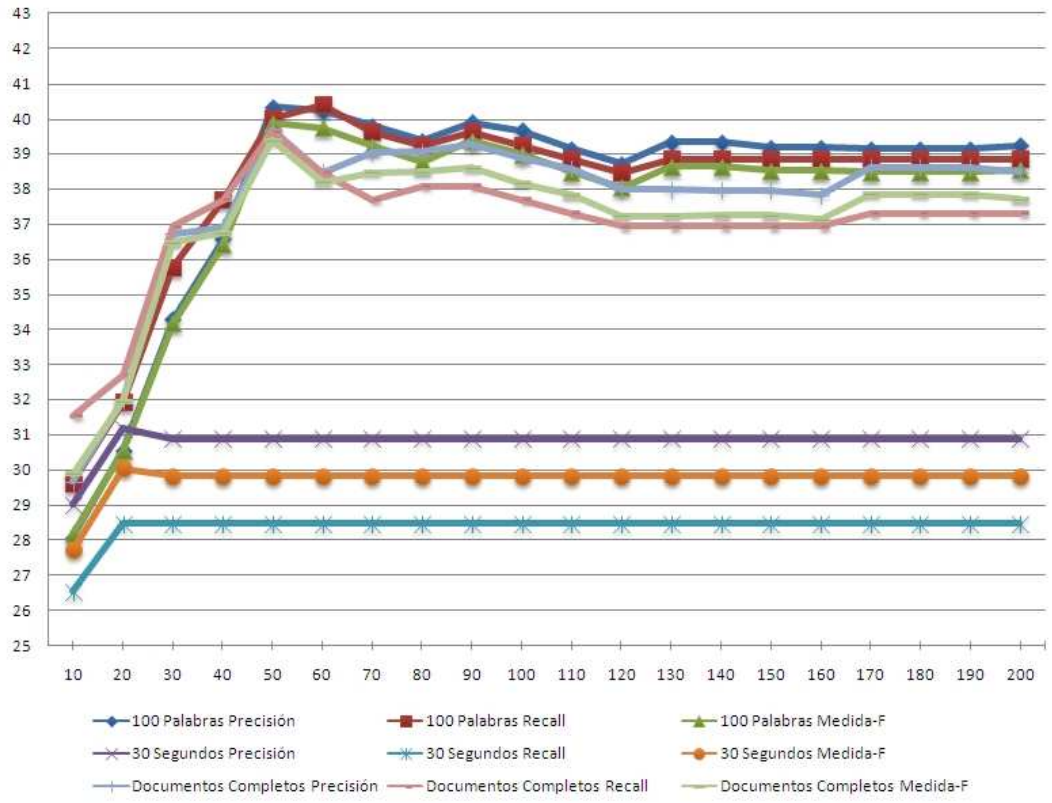
Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**3**	0	4	0	2	1	1	0	2	1	2	3	1
Cienc.	1	**16**	0	0	0	0	0	0	1	0	1	0	1
Cine	4	0	**5**	2	5	1	0	0	0	1	0	1	1
Come.	0	1	0	**6**	1	0	3	0	0	0	4	5	0
Dep	0	1	1	2	**5**	1	1	1	0	0	2	5	1
Educ.	1	0	2	2	0	**11**	0	0	0	0	3	1	0
Gent	4	0	2	4	0	1	**3**	1	0	0	1	2	2
Inst.	0	0	0	1	0	0	1	**13**	0	2	0	1	2
Jue.	0	0	0	0	0	0	1	0	**18**	0	1	0	0
Masc	1	4	1	0	0	1	0	0	1	**7**	3	0	2
Not.	0	2	1	3	1	1	2	0	1	0	**8**	0	1
Ocio	0	1	2	5	2	1	1	0	1	0	0	**3**	4
Viaj.	1	0	3	2	1	0	0	0	0	6	1	1	**5**

Tabla 22 - Matriz de Confusión - Clasificación basada en patrones - Documentos completos

Es fácil observar que el sistema confunde demasiado ciertas categorías. Es el caso de Viajes y Mascotas, que clasifica más documentos de viajes en la categoría Mascotas que en la propia. Al igual pasa con Ocio y Comedia, Gente con Automoción y con Comedia.

4.2.1.4 CONCLUSIONES SOBRE LA CLASIFICACIÓN BASADA EN PATRONES

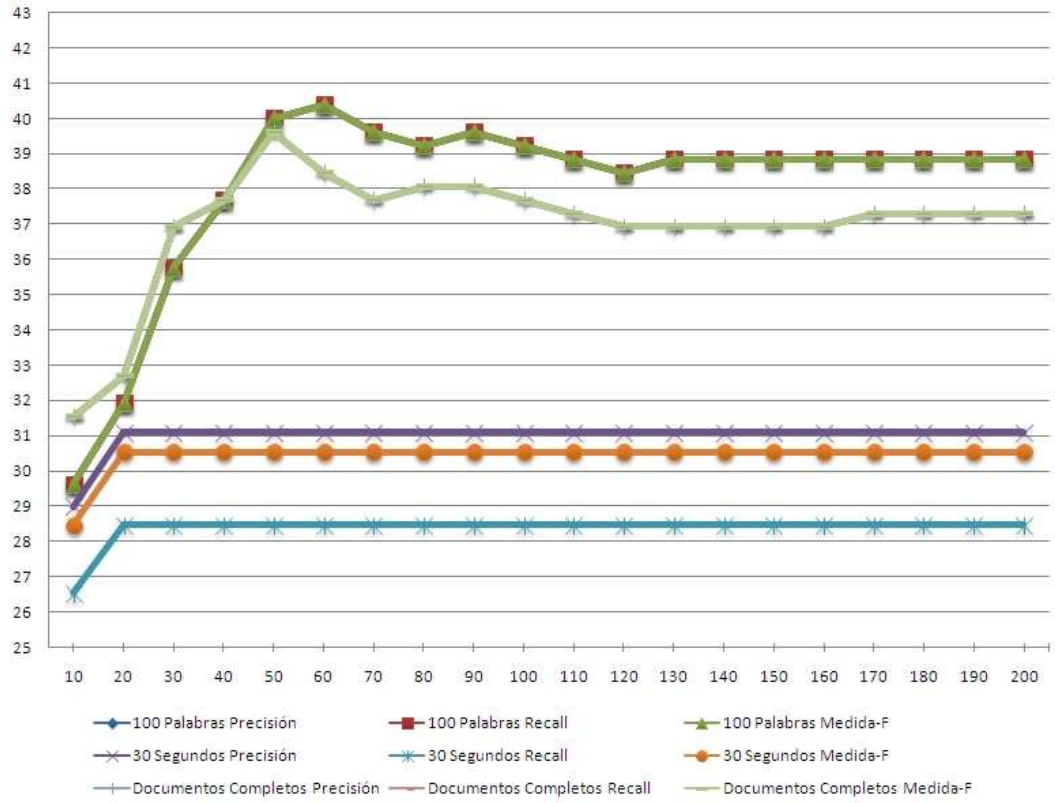
Para finalizar es preciso analizar la Gráfica 1, junto con todos los anteriores, y las dos siguientes gráficas donde se aprecia la relación entre las medidas de micro y macro averaging frente a la variación de palabras para cada tipo de entrenamiento.



Gráfica 5 - Clasificación Basada en patrones - Macro-Averaging

La característica más relevante en la medida de macro-averaging es el pico que marca la gráfica con la configuración del sistema de 50 y 60 palabras. La cobertura se mantiene muy ligada a la precisión en casi todos los esquemas analizados. Valores cercanos al 30% de cobertura indican que de los documentos que hay en una categoría únicamente clasifica bien el 30%, aunque la precisión esté por encima, esta baja cobertura hace del sistema un clasificador pobre.

Este efecto se aprecia también en la medida micro-averaging como muestra la siguiente tabla.



Gráfica 6 - Clasificación Basada en patrones - Micro-Averaging

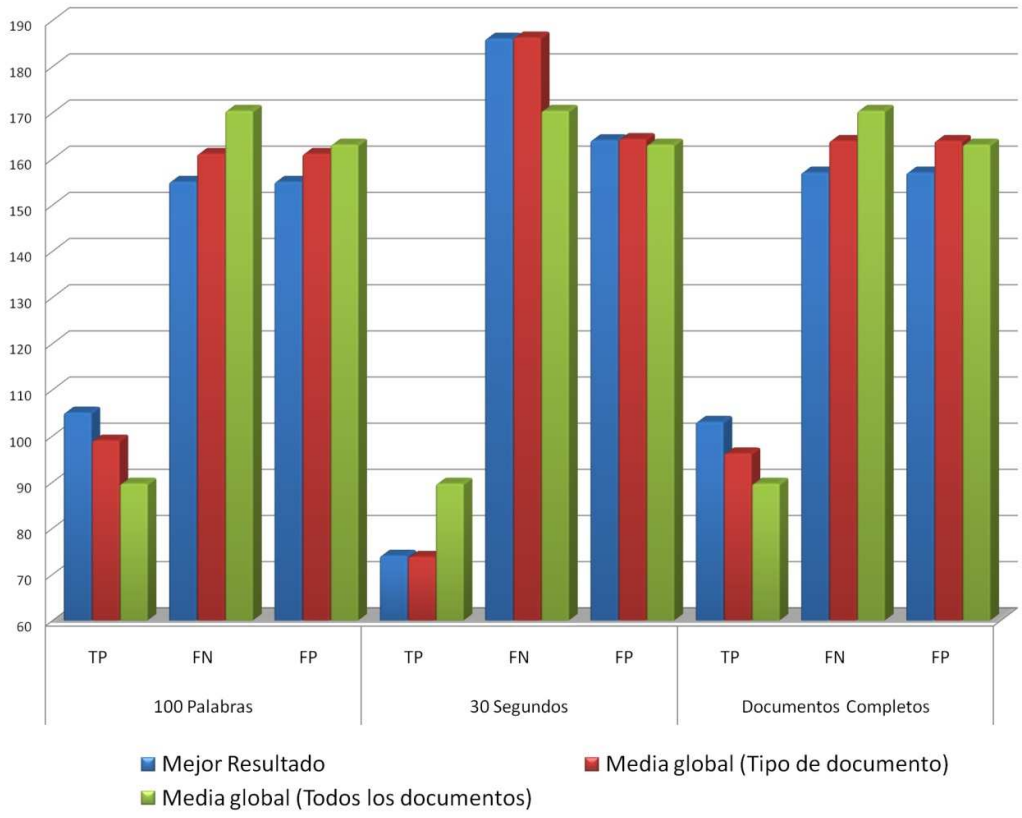
En cuanto a la comparación de las categorías, el siguiente diagrama de bloques compara el mejor resultado de cada prueba mejor valorada de los apartados anteriores junto con la media del sistema, hablando del número de FN, FP y TP.

El diagrama representa la mejor puntuación de cada tipo de documento, la media de las puntuaciones obtenidas en cada tipo de documento y la media de todas las pruebas realizadas en este tipo de clasificación, es decir, en la clasificación basada en patrones. Se ha escogido la media aritmética ya que es la que mejor representa a estos valores pues no existen valores demasiado atípicos en los resultados.

De manera un poco más detallada, se puede decir que la media global del sistema es la media de todos los resultados dados por el sistema, de esta manera se puede observar el comportamiento del sistema de manera global.

Por otro lado, la media global por tipo de documento hace referencia al comportamiento del sistema por tipo de documento.

Un sistema perfecto tendría el número de FP y FN a cero y el número de TP al máximo (el número de documentos del conjunto de test).



Gráfica 7 - Diagrama de Bloques - Clasificación Basada en Patrones - FP, FN, TP y medias

La única configuración que no da mejores resultados en todo con respecto a la media global de todos los documentos (color verde) es la configuración de los 30 primeros segundos de cada vídeo. Ni siquiera la mejor de las marcas supera los resultados de la media global. Este tipo de pruebas es sin duda la que peores resultados ha dado.

Aunque en el caso de las pruebas con documentos completos la media de las mismas no mejore la media global, por ejemplo en número de FP, es importante saber, que la calidad del sistema no se mide por un conjunto distinto de configuraciones, sino por la configuración concreta. Por tanto, para concluir, se puede decir que, las configuraciones del sistema, con casi iguales resultados, que obtienen una valoración aceptable son 100 palabras más repetidas y Documentos completos, siguiendo este mismo orden.

Por último, y escogiendo la mejor configuración del sistema, se van a analizar los resultados que se obtienen cuando se dan más de un resultado por documento. Los resultados están calculados desde un punto de vista global, ya que tanto el micro-averaging como el macro-averaging son calculables únicamente para conceptos binarios, es decir, aplicable a este sistema, para los aciertos, precisión, cobertura y otros en unos resultados obtenidos a la primera.

	P1	R1	F1
1	40,38	100	45,85
2	52,69	100	58,19
3	60,76	100	65,94
4	66,92	100	71,66
5	73,46	100	77,57
6	77,69	100	81,32
7	82,69	100	85,65
8	88,46	100	90,55
9	91,53	100	93,11
10	94,61	100	95,64
11	98,07	100	98,45
12	99,23	100	99,38
13	100	100	100

Tabla 23 - Clasificación basada en Patrones - Resultados finales

La probabilidad de acertar un documento es 1 entre el número de categorías, por lo tanto, en el caso del sistema tratado, se puede decir que es $^1/_{13} = 0{,}079$, es decir, existe un 8% de probabilidades de acertar. Si el sistema ofrece una precisión superior al 40%, se podría decir que es 5 veces superior a la probabilidad de acertar dando un resultado (Lift, ecuación 15), por lo tanto, los resultados obtenidos, se pueden decir que son bastante aceptables.

Por lo tanto, y para concluir, el sistema basado en patrones eligiendo el tipo de documento de las 100 palabras más repetidas y escogiendo la configuración de 60 palabras, es la que mayor efectividad ofrece.

4.2.2 CLASIFICACIÓN BASADA EN EJEMPLOS

La clasificación basada en ejemplos devuelve un conjunto de vídeos de las distintas categorías del sistema, se ponderan y se suman sus resultados de aquellos que pertenecen a la misma categoría.

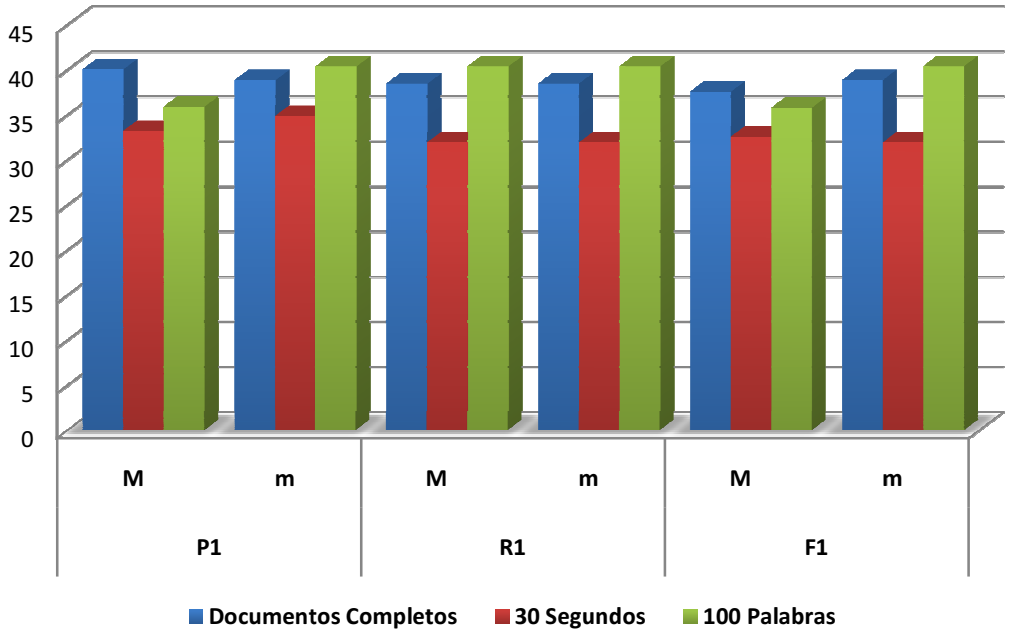
Los resultados obtenidos se muestran en la siguiente tabla (ver anexos para los datos completos).

Tipo de documento	Vídeos (vecinos)	Palabras	P1		R1		F1	
			M	m	M	m	M	m
Documentos Completos	45	40	40,05	38,85	38,46	38,46	37,57	38,85
	50	40	39,72	38,85	38,85	38,85	37,36	38,85
30 Segundos	50	90	33,2	34,87	31,92	31,92	32,57	31,92
	50	30	33,15	34,87	31,92	31,92	32,53	31,92
100 Palabras	**15**	**50**	**35,9**	**40,38**	**40,38**	**40,38**	**35,77**	**40,38**
	50	50	35,44	40	40	40	35,16	40

Tabla 24 -Mejores resultados - Clasificación basada en ejemplos (precisión en 1)

Para comprender bien las pruebas realizadas en este tipo de clasificación preciso recordar lo que son vídeos (entendido como vecinos) y palabras. Por palabras se entiende que son,

aquellas que el sistema considera más importantes, es decir, si un sistema tiene como configuración 30 palabras, esto significa que tomará las 30 palabras más importantes o que mejor representa a un documento. Por vídeos (vecinos) se entiende a la cantidad de vecinos más cercanos que devuelve el sistema (no olvidar que se trata de un sistema que implementa K-NN).



Gráfica 8 - Diagrama de bloques - Clasificación basada en ejemplos

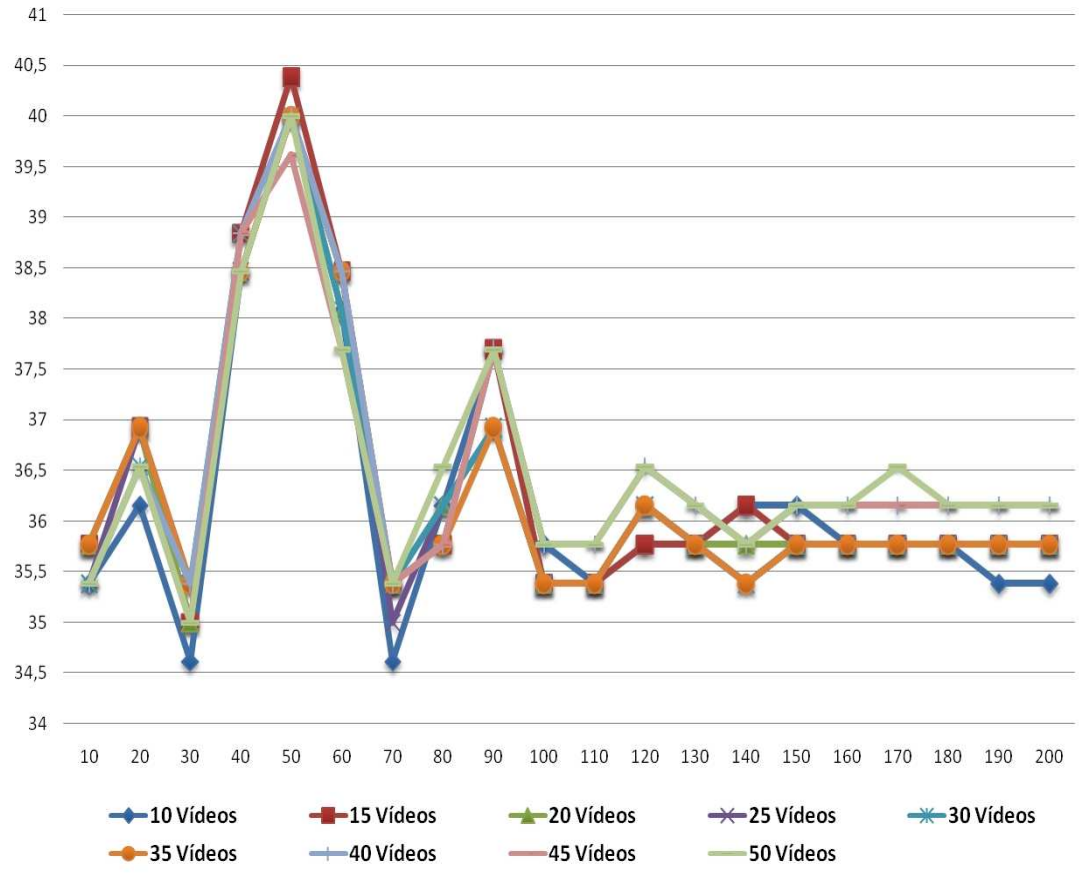
En este tipo de sistemas se aprecia un poco más la diferencia entre los tres tipos de documentos. Los documentos que únicamente contienen los 30 primeros segundos de cada vídeo obtienen peores resultados, mientras que aquellos, por el contrario, contienen todas las palabras del documento o, aquellas que más se repiten, obtienen mejores resultados.

Como en los casos anteriores, la cobertura global del sistema es bastante alta, y en particular, sigue siendo mayor a la precisión. En los siguientes apartados se van a detallar los resultados obtenidos con cada tipo de documento. No obstante, y como se ha comentado antes, estas pruebas se han realizado variando dos parámetros, por lo que los resultados obtenidos deben ser analizados en función de ambos, es decir, en función del número de palabras y en función del número de vídeos.

4.2.2.1 RESULTADOS DE LA CLASIFICACIÓN CON LAS CIEN PALABRAS MÁS IMPORTANTES

Los resultados obtenidos por esta clasificación están en los anexos. Puesto que la representación de estos datos en gráficas y tablas es bastante costosa, se ha preferido hacer constantemente referencias a los anexos y únicamente indicar en estos capítulos resúmenes o los mejores resultados.

La siguiente gráfica muestra la precisión del sistema, de manera global. La gráfica muestra el comportamiento cambiando el número de palabras, con respecto al número de vídeos.



Gráfica 9 - Clasificación Basada en Ejemplos - Variación de la precisión con respecto al número de palabras - 100 Palabras más repetidas

Se puede observar claramente, que el valor para el que la precisión es más alta es para aquellos, independientemente del número de vídeos, que poseen una configuración del sistema de 40 a 60 palabras.

Destaca entre todos, el valor que ofrece la clasificación que devuelve 15 vídeos y no parece haber una relación entre el número de vecinos (vídeos) y la precisión, en el sentido de que, a más vecinos más precisión.

A modo de resumen, la precisión para este tipo de sistemas, escogiendo el primer resultado (precisión en 1, P1), viene dado por la siguiente tabla:

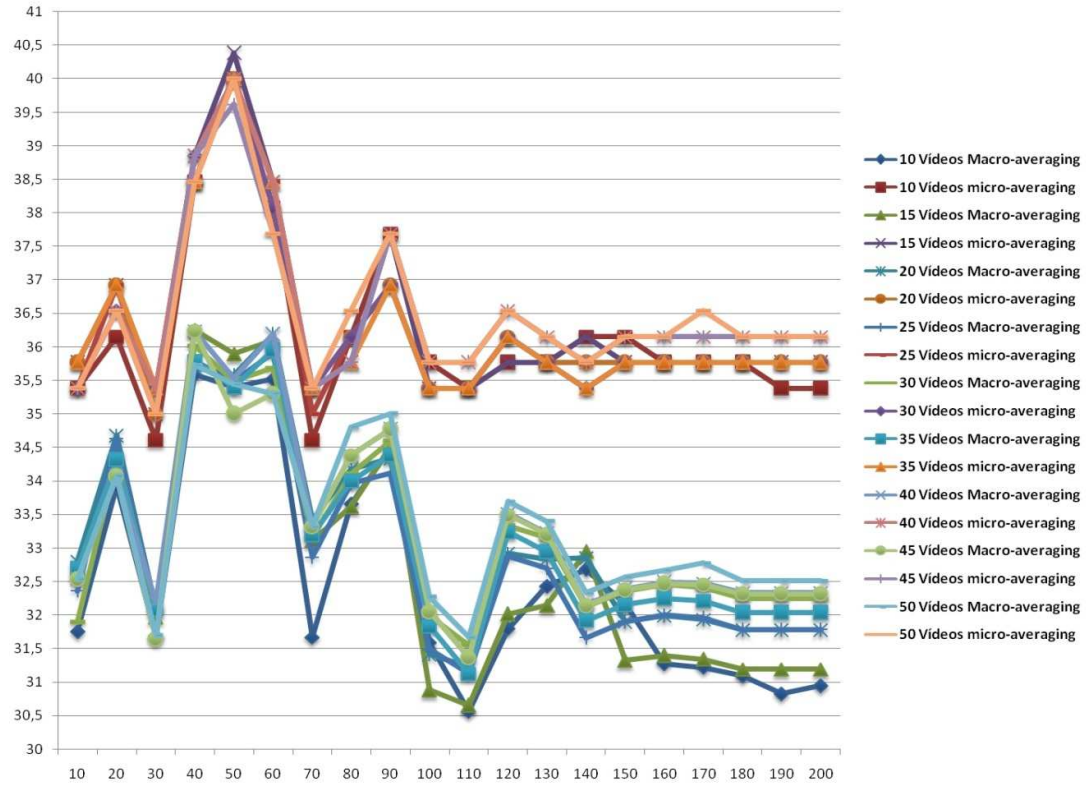
Palabras	Vecinos (Valor de K)								
	10	15	20	25	30	35	40	45	50
10	35,38	35,77	35,77	35,38	35,38	35,77	35,38	35,38	35,38
20	36,15	36,92	36,92	36,92	36,54	36,92	36,54	36,54	36,54
30	34,62	35	35	35,38	35,38	35,38	35,38	35	35
40	38,46	38,85	38,46	38,46	38,85	38,46	38,85	38,85	38,46
50	40	**40,38**	40	40	40	40	40	39,62	40
60	38,08	38,46	38,08	38,46	38,08	38,46	38,46	37,69	37,69
70	34,62	35,38	35,38	35	35,38	35,38	35,38	35,38	35,38
80	36,15	35,77	36,15	36,15	36,15	35,77	35,77	35,77	36,54
90	37,69	37,69	36,92	36,92	36,92	36,92	37,69	37,69	37,69
100	35,77	35,38	35,38	35,38	35,38	35,38	35,77	35,77	35,77
110	35,38	35,38	35,38	35,38	35,38	35,38	35,77	35,77	35,77
120	35,77	35,77	36,15	36,15	36,15	36,15	36,54	36,54	36,54
130	35,77	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15
140	36,15	36,15	35,77	35,38	35,38	35,38	35,77	35,77	35,77
150	36,15	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15
160	35,77	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15
170	35,77	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,54
180	35,77	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15
190	35,38	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15
200	35,38	35,77	35,77	35,77	35,77	35,77	36,15	36,15	36,15

Tabla 25 - Clasificación Basada en Ejemplos – Precisión (P1) - 100 Palabras más repetidas

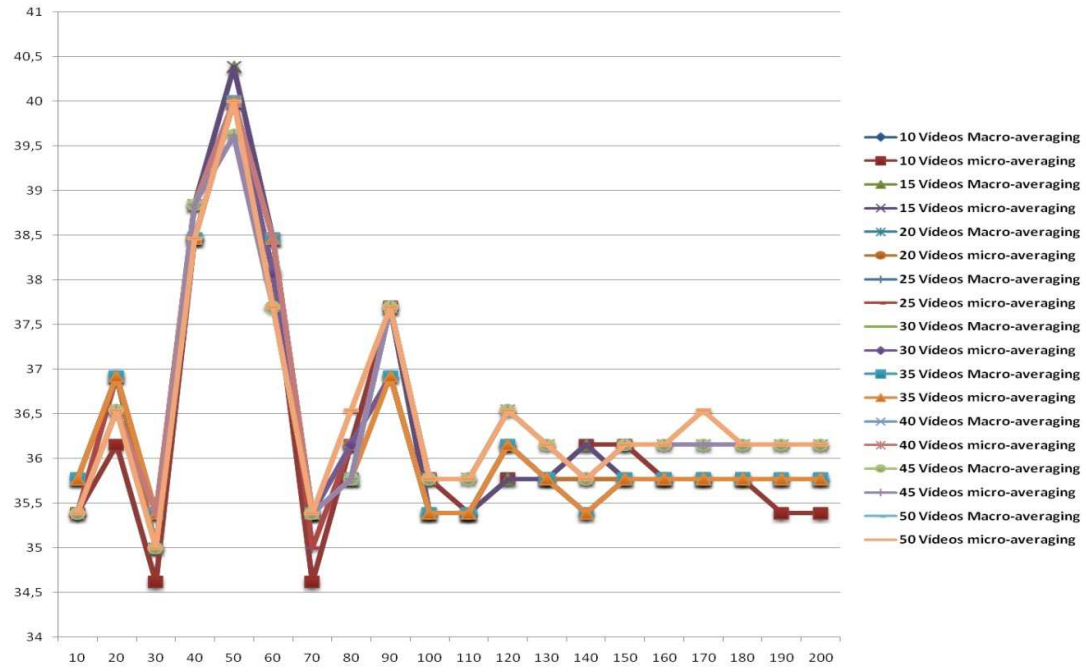
Centrándose ahora en lo que ofrece el análisis de las categorías, las siguientes gráficas muestran la cobertura y la precisión del sistema, con respecto al número de vídeos y palabras.

En la Gráfica 10, se observa la diferencia entre el macro y el micro averaging en cuanto a la precisión, al igual que en los términos globales, los valores que más destacan, aquellas configuraciones que mejor resultados dan son las que comprenden valores para las palabras de entre 40 y 50. Es el caso también de la cobertura, como se muestra en la Gráfica 11. En la precisión se nota una ligera diferencia entre el micro-averaging y el macro-averaging mientras que en la cobertura del sistema se mantienen iguales o similares.

No obstante, y como se ha comentado, ni la cobertura ni la precisión parecen depender del número de vecinos (vídeos), no hay una relación entre ellos, da igual el número de vídeos, lo más importante y el factor que marca la diferencia entre unos clasificadores y otros es el número de palabras analizadas.



Gráfica 10 - Clasificación basada en ejemplos – Precisión (P1) - Micro y macro averaging - 100 Palabras más repetidas



Gráfica 11 - Clasificación basada en ejemplos - Cobertura - Micro y macro averaging - 100 Palabras más repetidas

Centrándose en la configuración del sistema en aquella que da como resultado 15 vídeos y utiliza 50 palabras, las medidas obtenidas para las categorías son las mostradas en la tabla.

Categorías	FN	FP	TP	P1	R1	F1
Automoción	16	10	4	28,57	20	26,32
Ciencia y tecnología	0	20	20	50	100	55,56
Cine y animación	15	10	5	33,33	25	31,25
Comedia	15	14	5	26,32	25	26,04
Deportes	17	8	3	27,27	15	23,44
Educación	10	16	10	38,46	50	40,32
Gente y blogs	18	19	2	9,524	10	9,615
Instrucc. varias y estilo	4	11	16	59,26	80	62,5
Juegos	2	12	18	60	90	64,29
Mascotas y animales	**12**	**5**	**8**	**61,54**	**40**	**55,56**
Noticias y política	8	11	12	52,17	60	53,57
Ocio	19	7	1	12,5	5	9,615
Viajes y eventos	19	12	1	7,692	5	6,944
Total	155	155	105	-	-	-
Macro-Averaging				**35,9**	**40,38**	**35,77**
Micro-Averaging				**40,38**	**40,38**	**40,38**

Tabla 26 - Características de la evaluación por categorías - Clasificación basada en ejemplos - 100 Palabras más repetidas

Lo que más llama la atención de esta configuración es la diferencia tan alta entre unas categorías y otras, por lo general, aquellas que tienen una buena precisión tienen una buena cobertura, y a su vez son relativamente altas. Por el contrario, los que tienen mala precisión o cobertura, poseen valores bastante bajos. Otro dato importante es la categoría de Ciencia y Tecnología, que, aun clasificando todos sus vídeos correctamente comete tantos fallos como aciertos, de ahí su alta cobertura y su baja precisión.

La matriz de confusión para este sistema es la mostrada en la tabla.

Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**4**	3	0	0	1	2	4	0	2	1	1	0	2
Cienc.	0	**20**	0	0	0	0	0	0	0	0	0	0	0
Cine	3	1	**5**	2	1	0	2	2	1	1	1	0	1
Come.	1	1	0	**5**	1	1	1	3	0	0	4	1	2
Dep	0	1	4	3	**3**	3	3	1	0	0	1	1	0
Educ.	1	3	1	0	0	**10**	0	1	1	0	1	1	1
Gent	1	2	0	4	0	4	**2**	1	3	0	2	1	0
Inst.	0	0	0	1	1	0	0	**16**	0	1	0	0	1
Jue.	0	0	0	0	0	0	1	0	**18**	0	0	1	0
Masc	0	5	0	0	0	2	1	1	2	**8**	0	0	1
Not.	0	0	0	0	0	0	2	1	1	0	**12**	2	2
Ocio	1	2	1	4	1	2	5	0	1	0	0	**1**	2
Viaj.	3	2	4	0	3	2	0	1	1	2	1	0	**1**

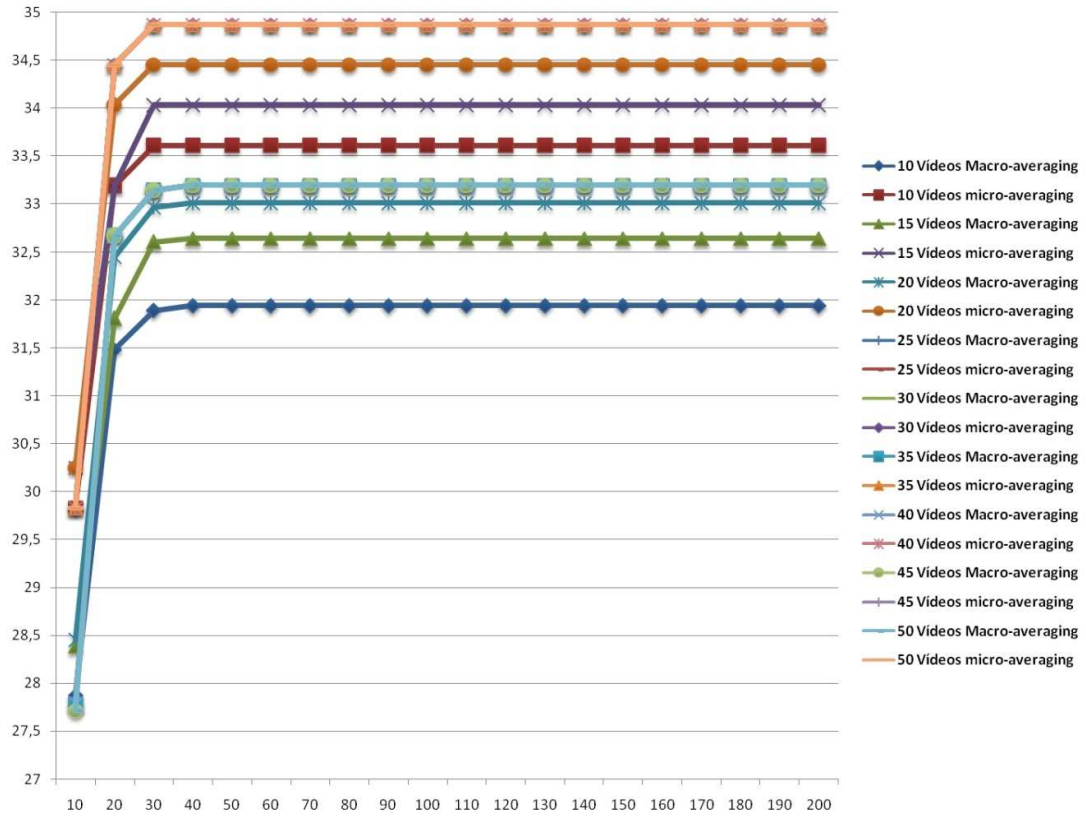
Tabla 27 - Matriz de Confusión - Clasificación basada en ejemplos - 100 palabras más repetidas

Como se ha comentado y como se aprecia aquí mejor, el sistema ha clasificado 40 vídeos en la categoría Ciencia de los cuales 20 son erróneos y otros 20 no, lo que quiere decir, que el sistema identifica más el 15% de los vídeos en Ciencia, y sobre todo confunde aquellos que son de Mascotas.

A destacar también están las categorías de Ocio y Viajes, ya que el sistema confunde casi todos sus vídeos con el resto de categorías, de hecho, la precisión de estas categorías ronda el 10% y la cobertura apenas un 5%.

4.2.2.2 RESULTADOS DE LA CLASIFICACIÓN CON LOS TREINTA PRIMEROS SEGUNDOS

La siguiente gráfica muestra el comportamiento de los clasificadores usados con este tipo de documentos, mostrando la precisión al variar el número de vecinos (vídeos) y de palabras. Como se ve, parece que al aumentar el número de vecinos aumenta la precisión, al igual que con el número de palabras, pero, llega un momento en el que la precisión se estanca, sin bajar ni subir de valor. De igual manera se comporta la cobertura y la medida-F.



Gráfica 12 - Clasificación basada en ejemplos - Precisión total - 30 primeros segundos

Ocurre, como es lógico, lo mismo con los cálculos de macro y micro averaging para cada una de las medidas anteriores. Por lo tanto, es más importante analizar aquella configuración que se ha escogido (50 vídeos, 90 palabras) en detalle.

Categorías	FN	FP	TP	P1	R1	F1
Automoción	15	12	5	29,41	25	28,41
Ciencia y tecnología	9	10	11	52,38	55	52,88
Cine y animación	16	20	4	16,67	20	17,24
Comedia	18	10	2	16,67	10	14,71
Deportes	19	14	1	6,667	5	6,25
Educación	13	12	7	36,84	35	36,46
Gente y blogs	18	7	2	22,22	10	17,86
Instrucc. varias y estilo	6	6	14	70	70	70
Juegos	**3**	**7**	**17**	**70,83**	**85**	**73,28**
Mascotas y animales	18	19	2	9,524	10	9,615
Noticias y política	12	17	8	32	40	33,33
Ocio	16	14	4	22,22	20	21,74
Viajes y eventos	14	7	6	46,15	30	41,67
Total	177	155	83	-	-	-
			Macro-Averaging	**33,2**	**31,92**	**32,57**
			Micro-Averaging	**34,87**	**31,92**	**31,92**

Tabla 28 - Características de la evaluación por categorías - Clasificación basada en ejemplos - 30 primeros segundos

Una vez más, la categoría Juegos es la que más precisión da, y en este caso también es la que mayor cobertura ofrece. El resto de categorías ofrecen valores muy distintos, pero entre ellas destaca la categoría Deportes que, como se observa, únicamente da un acierto y no solo eso, sino que además, da 14 fallos, con lo que, ni su precisión ni su cobertura alcanzan el 10%.

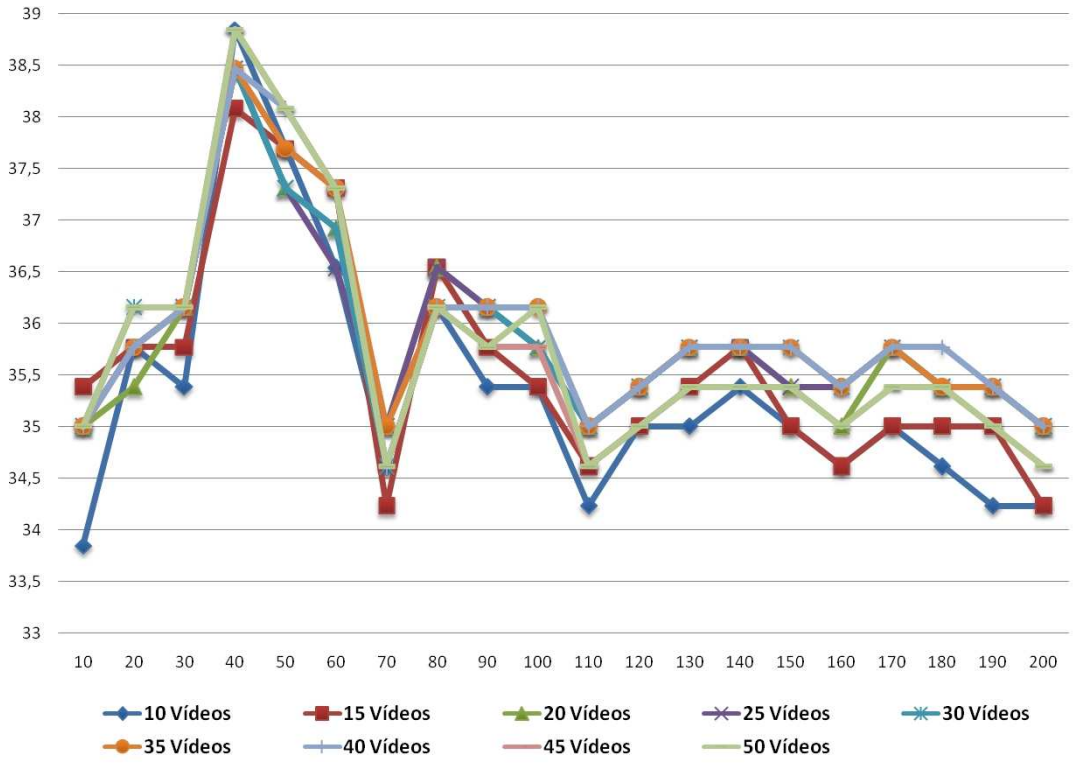
Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**5**	1	2	0	3	1	2	0	0	3	1	0	1
Cienc.	1	**11**	2	0	0	2	0	1	0	0	1	1	0
Cine	3	0	**4**	0	0	0	0	1	2	0	1	2	1
Come.	0	2	2	**2**	2	2	0	0	1	4	2	2	0
Dep	0	0	0	0	**1**	1	2	2	1	2	2	2	0
Educ.	1	1	4	0	3	**7**	0	0	1	2	1	0	0
Gent	3	2	2	3	0	0	**2**	0	1	3	1	1	1
Inst.	0	0	1	2	1	0	1	**14**	0	0	1	0	0
Jue.	0	0	0	0	0	1	0	0	**17**	0	1	0	1
Masc	2	1	3	1	0	0	1	1	0	**2**	3	1	2
Not.	0	0	1	1	0	3	0	1	1	1	**8**	3	0
Ocio	1	1	2	1	4	1	1	0	0	2	2	**4**	1
Viaj.	1	2	1	2	1	1	0	0	0	2	1	2	**6**

Tabla 29 - Matriz de Confusión - Clasificación basada en ejemplos - 30 primeros segundos

La categoría que más llama la atención es Deportes. Sus vídeos, además de no haber sido clasificados 7 de ellos en ninguna categoría, confunde las categorías Automoción, Comedia y Ocio con ella. En el caso de la categoría Instrucciones, se observa claramente que no confunde demasiado dicha categoría con otras, pues de 20 vídeos que clasifica, el 70% (14 vídeos) lo hace en la dicha y el resto, únicamente 6 vídeos de otras categorías, decide que son de ella.

4.2.2.3 RESULTADOS DE LA CLASIFICACIÓN CON EL DOCUMENTO COMPLETO

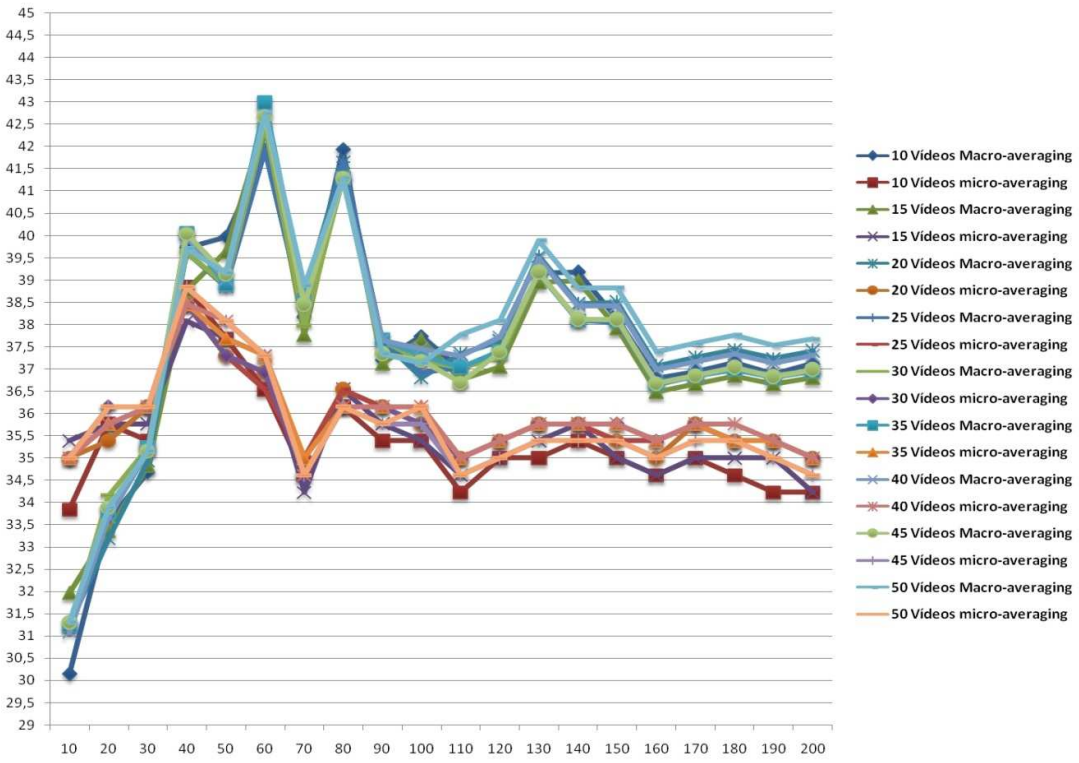
Por último, se va a analizar la clasificación de documentos completos siguiendo el mismo esquema que el seguido hasta ahora. La precisión, la cobertura y la medida-F global de estos sistemas vienen representadas en las tablas de los anexos. No obstante, de manera esquemática y puesto que la precisión parece que es la medida que más caracteriza a estos sistemas, la siguiente gráfica muestra las distintas precisiones del sistema.



Gráfica 13 - Clasificación basada en Ejemplos - Precisión (P1) - Documentos completos

El dato más relevante, además del pico que hace la precisión en el punto en el que se tienen 40 palabras, es que, una vez alcanzado este punto, a medida que se van aumentando el número de palabras, la precisión disminuye. La cobertura, no obstante, y como en el resto de sistemas comentados y estudiados, se mantiene al 100% al clasificar todos los documentos del conjunto de test.

Es importante observar, en la gráfica Gráfica 14, la subida que hacen los valores cuando rondan las 60 y las 80 palabras y sobre todo comparar esto con la precisión global y la precisión micro-averaging. No obstante, y como en casos anteriores y ya comentado, se observa fácilmente que la precisión no depende del número de vídeos dados, sino del número de palabras, pues las líneas que representan cada variación de vídeos, parece que van unidas.



Gráfica 14 - Clasificación basada en Ejemplos - Precisión Macro-averaging y micro-averaging - Documentos completos

Por lo tanto, y escogiendo el clasificador con la configuración de 45 vídeos y 40 palabras, los resultados obtenidos para las categorías son los mostrados en la tabla.

Categorías	FN	FP	TP	P1	R1	F1
Automoción	15	3	5	62,5	25	48,08
Ciencia y tecnología	1	21	19	47,5	95	52,78
Cine y animación	15	11	5	31,25	25	29,76
Comedia	18	14	2	12,5	10	11,9
Deportes	17	4	3	42,86	15	31,25
Educación	10	15	10	40	50	41,67
Gente y blogs	17	27	3	10	15	10,42
Instrucc. varias y estilo	5	28	15	34,88	75	39,06
Juegos	**3**	**3**	**17**	**85**	**85**	**83,33**
Mascotas y animales	12	4	8	66,67	40	58,82
Noticias y política	10	13	10	43,48	50	44,64
Ocio	18	5	2	28,57	10	20,83
Viajes y eventos	18	11	2	15,38	10	13,89
Total	159	159	101	-	-	-
Macro-Averaging				**40,05**	**38,85**	**37,42**
Micro-Averaging				**38,85**	**38,85**	**38,46**

Tabla 30 - Características de la evaluación por categorías - Clasificación basada en ejemplos - Documentos Complemtos

En este caso, una de las categorías que mayor precisión tiene es Mascotas al igual que Automoción, pero es importante analizar ambas categorías, ya que, como se observa, la cobertura que presentan es muy baja. Si el sistema se comportase acorde a estas dos categorías, el sistema tendría una buena precisión, clasificaría los documentos correctamente, pero tendría una pésima cobertura, lo que daría como resultado un sistema con pocos fallos pero con muy pocos resultados dados, es decir, muy fiable pero poco exhaustivo. La matriz de confusión ayuda a analizarlo mejor.

Real / Predicha	Aut.	Cienc.	Cine	Come.	Dep	Educ.	Gent.	Inst.	Jue.	Masc.	Not.	Ocio	Viaj.
Aut.	**5**	2	0	0	1	1	7	1	0	1	1	0	1
Cienc.	0	**19**	0	0	0	0	0	1	0	0	0	0	0
Cine	1	0	**5**	2	1	1	4	5	0	0	0	0	1
Come.	0	4	0	**2**	1	1	2	4	0	0	5	1	0
Dep	0	0	3	1	**3**	2	6	4	0	0	0	0	1
Educ.	0	3	1	1	0	**10**	0	0	0	0	3	1	1
Gent	0	4	2	4	1	3	**3**	0	0	0	2	1	0
Inst.	0	0	1	2	0	0	0	**15**	1	0	0	1	0
Jue.	0	0	0	0	0	0	0	1	**17**	1	0	0	1
Masc	0	4	0	0	0	2	1	4	0	**8**	0	0	1
Not.	0	1	2	1	0	0	1	3	0	0	**10**	1	1
Ocio	0	2	0	3	0	2	5	1	1	0	0	**2**	4
Viaj.	2	1	2	0	0	3	1	4	1	2	2	0	**2**

Tabla 31 - Matriz de Confusión - Clasificación basada en ejemplos - Documentos Completos

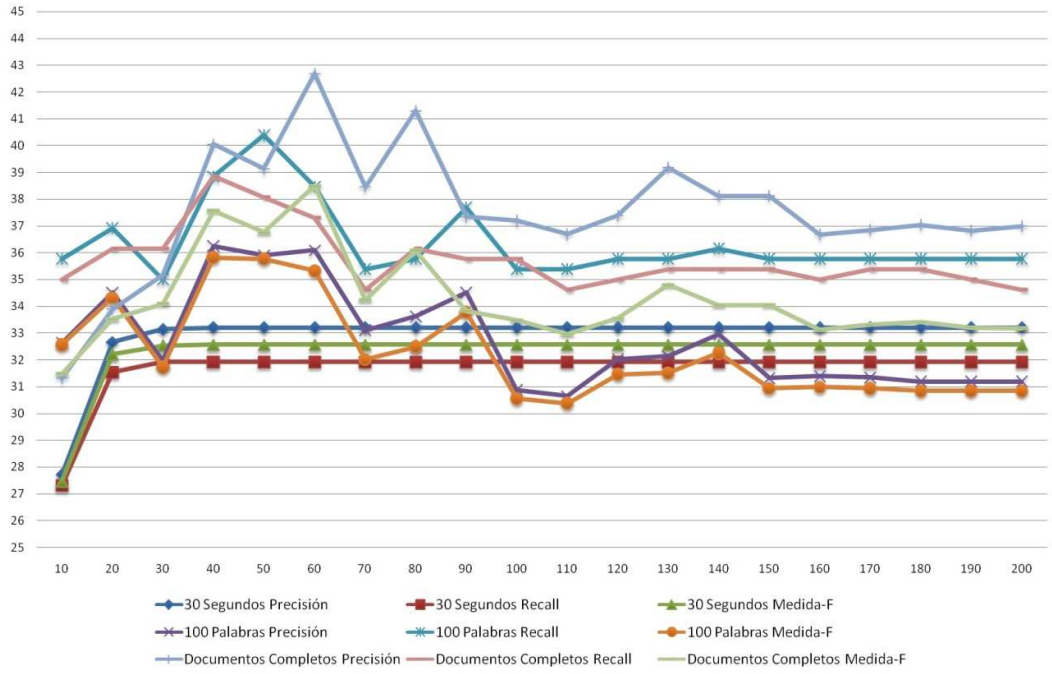
Como se ha comentado antes, y en la tabla anterior se aprecia mejor, se observa que la categoría Mascotas y la categoría automoción son de las pocas en las que apenas el sistema ha confundido los documentos de otras categorías con ella. No obstante, su cobertura es poca ya que confundió muchos de sus vídeos con otras categorías.

El caso contrario es el caso de la categoría Ciencia, sus vídeos, han sido clasificados, en la gran mayoría bien, pero el sistema ha confundido muchos de los vídeos de otras categorías con ésta. De ahí su alta cobertura y su baja precisión.

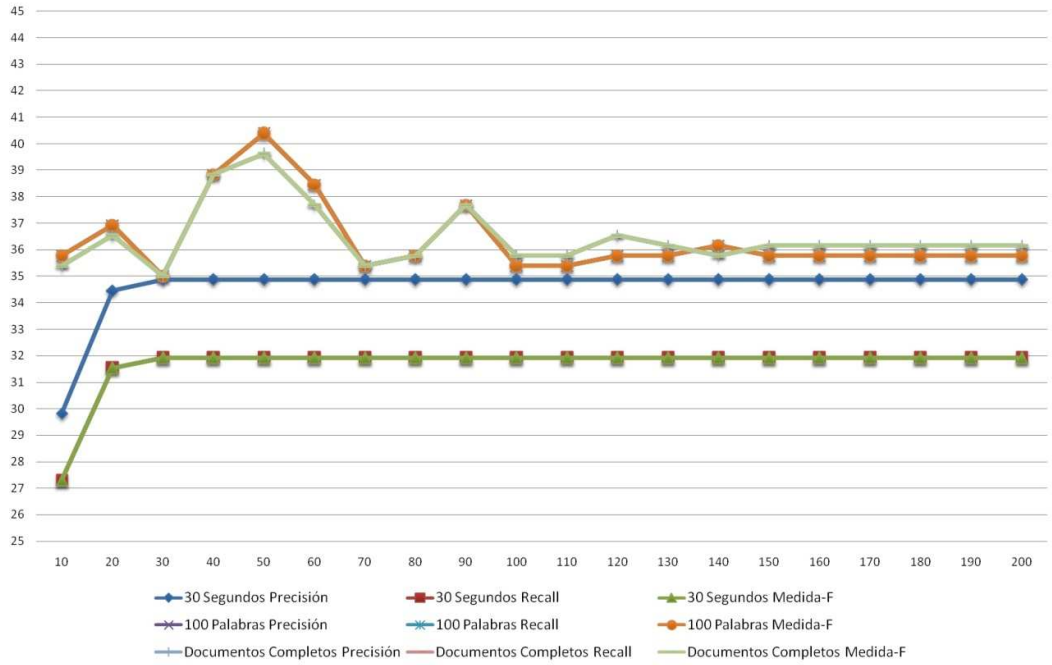
Un buen ejemplo de un clasificador eficaz sería aquél que se comportase en todas las categorías como en el caso de la categoría Juegos, donde clasifica la gran mayoría de sus vídeos bien y no confunde casi ningún vídeos de otras categorías con ella.

4.2.2.4 CONCLUSIONES SOBRE LA CLASIFICACIÓN BASADA EN EJEMPLOS

Una vez terminado el estudio hasta aquí del sistema basado en ejemplos, es necesario unir todos los resultados para sacar una conclusión de los mismos. Primeramente es preciso volver a la Tabla 24 donde se puede observar un resumen general de los resultados obtenidos en este tipo de clasificación y dependiendo, tanto del número de vecinos como del número de palabras. Centrándose entonces en los que mejores resultados han dado, y viendo una comparación de las correspondientes precisión, cobertura y medida-F con los diferentes cálculos de micro y macro averaging, se obtiene lo siguiente.



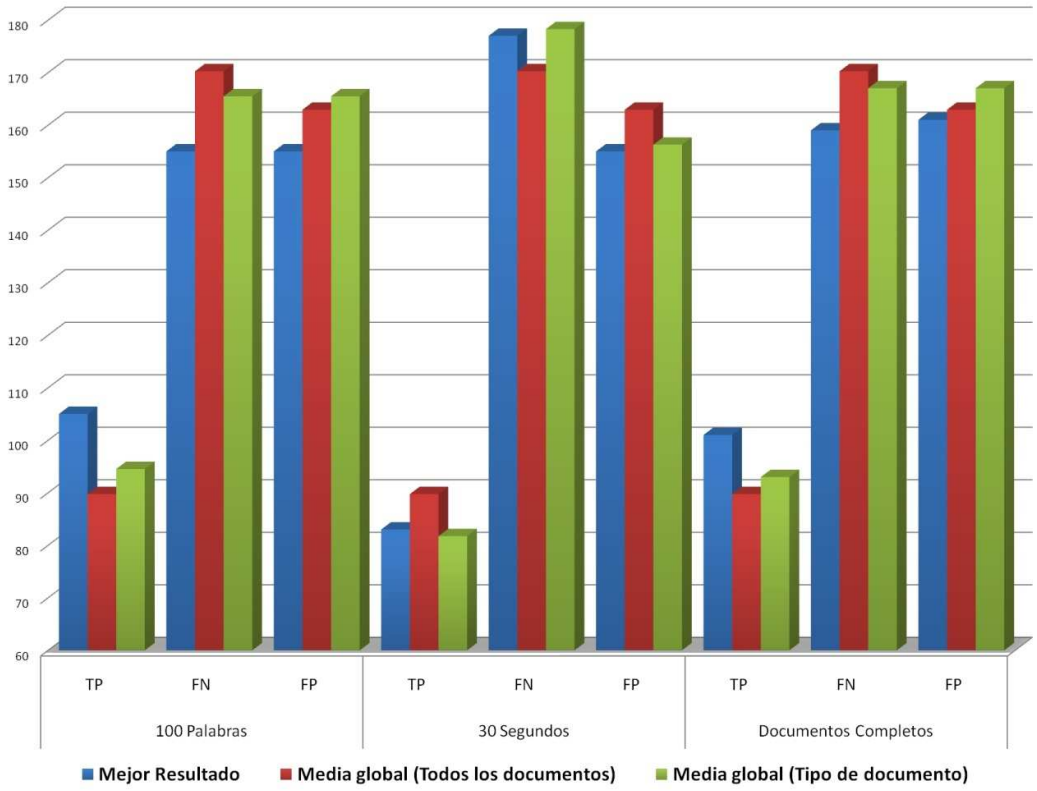
Gráfica 15 - Clasificación basada en ejemplos - Macro-averaging



Gráfica 16 - Clasificación basada en ejemplos - Micro-averaging

Si se observan los datos de la tablas correspondientes a este tipo de clasificación en los anexos y en las anteriores gráficas, se puede observar la diferencia entre unos sistemas y otros, dependiendo el tipo de documento. Los documentos completos y aquellos que contienen las cien palabras más repetidas, devuelven unos resultados bastante similares, y se diferencian en

pequeñas variaciones. No obstante, a medida que se aumentan el número de palabras, los resultados tienden a empeorar y a asemejarse a aquellos sistemas que hacen la clasificación con los 30 primeros segundos de cada vídeo. Incluso, en el caso de la cobertura y para el cálculo de macro-averaging llegan a ser peores que en el caso de los documentos incompletos (30 segundos). Valorando ahora las categorías y haciendo un análisis de los correspondientes FP, TP y FN, se obtiene el siguiente diagrama de bloques:



Gráfica 17 - Diagrama de Bloques - Clasificación Basada en Ejemplos - FP, FN, TP y medias

El anterior diagrama representa cada uno de los mejores resultados obtenidos con los distintos tipos de documentos, comparándolos con la media global de cada tipo de documento y con la media global en conjunto. Rápidamente se observa que los sistemas basados en los 30 primeros segundos de cada documento dejan un resultado bastante peor que el resto llegando incluso a empeorar la media global el mejor de sus resultados.

Para hacer una comparativa correcta y analizar los resultados correctamente, hay que tener en cuenta, que una clasificación perfecta, tendría el número de TP a 260, es decir, el máximo número de documentos, y tanto FP como FN a cero.

El resultado obtenido por documentos completos y por los documentos que contienen las cien palabras más repetidas, obtienen unos resultados que mejoran la media global incluso en lo que se refiere a la media de los documentos.

Por lo tanto, y analizando los resultados de los anteriores apartados, se puede concluir con que el mejor sistema basado en ejemplos es aquel que tiene los documentos con las cien palabras más repetidas, con la configuración de 15 vecinos y 50 palabras.Para dicha configuración la tabla, en términos globales, que resume esta información es la siguiente:

	P1	R1	F1
1	40,38	100	45,45
2	47,30	100	52,49
3	54,61	100	59,69
4	64,23	100	68,10
5	71,15	100	74,11
6	77,30	100	78,94
7	80,76	100	81,32
8	83,84	100	81,65
9	85	100	81,65
10	86,15	100	81,65
11	86,53	100	81,65
12	86,53	100	81,65
13	86,53	100	81,65

Tabla 32 - Clasificación basada en Ejemplos - Resultados finales

En este caso, al dar la misma precisión, se puede terminar con las mismas conclusiones ofrecidas en la clasificación basada en patrones. El sistema ofrece una precisión superior al 40%, se podría decir que es 5 veces superior a la probabilidad de acertar dando un resultado.

4.3 COMPARATIVA DE RESULTADOS

Para hacer una comparativa de todos los resultados, se han escogido aquellos que se decidieron en su momento que eran los mejores y se han combinado, para observar y analizar todas juntas y decidir así que configuración y qué tipo de sistema es el más eficiente y adecuado.

Para hacer la comparativa global, lo primero es visualizar en una tabla los mejores resultados escogidos hasta ahora(Tabla 33).

Tipo	Tipo de Documento	Vid.	Pal.	P1		R1		F1	
				M	m	M	m	M	m
Ejemplos	Doc. Com.	45	40	40,05	38,85	38,46	38,46	37,57	38,85
	30 Seg.	50	90	33,2	34,87	31,92	31,92	32,57	31,92
	100 Pal.	15	50	35,9	40,38	40,38	40,38	35,77	40,38
Patrones	Doc. Com.	-	50	39,68	39,62	39,615	39,615	39,36	39,62
	30 Seg.	-	60	30,88	31,09	28,46	28,46	29,81	30,53
	100 Pal.	**-**	**60**	**40,21**	**40,38**	**40,38**	**40,38**	**39,73**	**40,38**

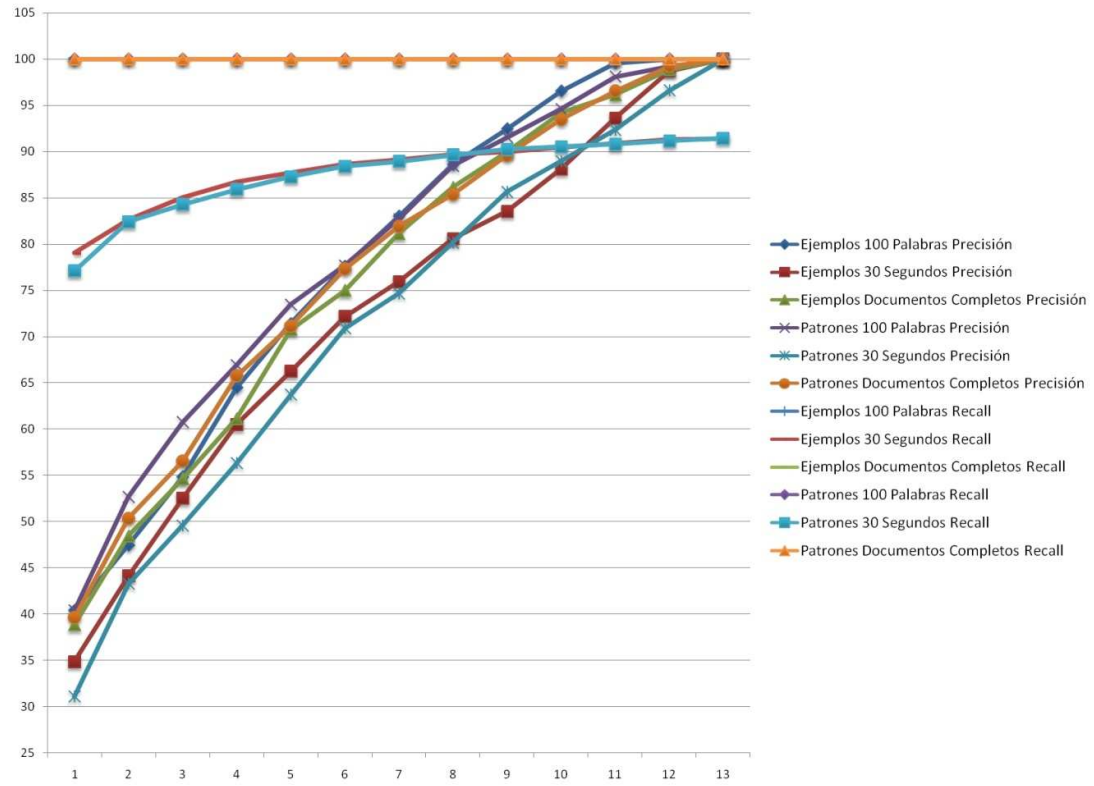
Tabla 33 - Resumen de mejores resultados

Los datos globales de las mejores pruebas quedan de la siguiente manera, donde se representa la precisión, la cobertura y la medida-F con respecto al número de resultados (desde 1 hasta 13):

Tipo de Clasificador			1	2	3	4	5	6	7	8	9	10	11	12	13
Ej.	100 P.	P	40,3	47,4	54,8	64,4	71,4	77,6	83	88,6	92,4	96,5	99,5	100	100
		R	100	100	100	100	100	100	100	100	100	100	100	100	100
		F	45,8	53,0	60,2	69,4	75,7	81,2	85,9	90,6	93,8	97,2	99,6	100	100
	30 S.	P	34,8	44,1	52,5	60,5	66,2	72,1	75,9	80,5	83,5	88,1	93,6	98,7	100
		R	79,0	82,6	85,0	86,7	87,7	88,6	89,1	89,6	90	90,4	90,9	91,3	91,4
		F	39,2	48,6	56,8	64,4	69,6	74,9	78,2	82,2	84,7	88,5	93,0	97,1	98,1
	D.C.	P	38,8	48,4	54,6	61,1	70,7	75	81,1	86,1	90	94,2	96,1	98,8	100
		R	100	100	100	100	100	100	100	100	100	100	100	100	100
		F	44,2	54,0	60,0	66,3	75,1	78,9	84,3	88,6	91,8	95,3	96,9	99,0	100
Pa.	**100 P.**	**P**	**40,3**	**52,6**	**60,7**	**66,9**	**73,4**	**77,6**	**82,6**	**88,4**	**91,5**	**94,6**	**98,0**	**99,2**	**100**
		R	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**	**100**
		F	**45,8**	**58,2**	**65,9**	**71,6**	**77,5**	**81,3**	**85,6**	**90,5**	**93,1**	**95,6**	**98,4**	**99,3**	**100**
	30 S.	P	31,0	43,2	49,5	56,3	63,7	70,8	74,6	80,1	85,6	88,9	92,3	96,6	100
		R	77,0	82,4	84,2	85,9	87,2	88,4	88,9	89,6	90,2	90,5	90,7	91,1	91,4
		F	35,3	47,8	54,0	60,4	67,3	73,8	77,1	81,9	86,5	89,2	92,0	95,4	98,1
	D.C.	P	39,6	50,3	56,5	65,7	71,1	77,3	81,9	85,3	89,6	93,4	96,5	99,2	100
		R	100	100	100	100	100	100	100	100	100	100	100	100	100
		F	45,0	55,9	61,9	70,6	75,5	80,9	85	87,9	91,5	94,7	97,2	99,3	100

Tabla 34 - Resultados Finales

De manera un poco más representativa se obtiene la gráfica de la anterior tabla:



Gráfica 18 - Resultados Finales

La precisión, tanto en el caso de la clasificación basada en ejemplos como en el caso de la clasificación basada en patrones sigue, más o menos, una misma función creciente con respecto al

número de resultados dados, la diferencia entre unos sistemas y otros es el punto de inicio y la precisión al dar un único resultado.

Categoría	Basado en Patrones			Basado en Ejemplos		
	100 Pal.	30 Seg.	Doc. Comp.	100 Pal.	30 Seg.	Doc. Comp.
Automoción	18,75	46,67	20	28,57	29,41	62,5
Ciencia y tecnología	61,54	54,17	64	50	52,38	47,5
Cine y animación	19,05	25	23,81	33,33	16,67	31,25
Comedia	24	13,33	22,22	26,32	16,67	12,5
Deportes	21,74	5,263	29,41	27,27	6,667	42,86
Educación	61,11	22,58	61,11	38,46	36,84	40
Gente y blogs	25	12,5	23,08	9,524	22,22	10
Instrucc. varias y estilo	82,35	73,33	86,67	59,26	70	34,88
Juegos	85,71	52,38	75	60	70,83	85
Mascotas y animales	40	20	41,18	61,54	9,524	66,67
Noticias y política	37,5	40,91	30,77	52,17	32	43,48
Ocio	23,81	25	13,64	12,5	22,22	28,57
Viajes y eventos	22,22	10,34	25	7,692	46,15	15,38
Macro-Averaging	**40,21**	**30,88**	**39,68**	**35,9**	**33,2**	**40,05**
Micro-Averaging	**40,38**	**31,09**	**39,62**	**40,38**	**34,87**	**38,85**

Tabla 35 - Precisión en categorías - Resultados Finales

Categoría	Basado en Patrones			Basado en Ejemplos		
	100 Pal.	30 Seg.	Doc. Comp.	100 Pal.	30 Seg.	Doc. Comp.
Automoción	15	35	15	20	25	25
Ciencia y tecnología	80	65	80	100	55	95
Cine y animación	20	10	25	25	20	25
Comedia	30	10	30	25	10	10
Deportes	25	5	25	15	5	15
Educación	55	35	55	50	35	50
Gente y blogs	15	5	15	10	10	15
Instrucc. varias y estilo	70	55	65	80	70	75
Juegos	90	55	90	90	85	85
Mascotas y animales	20	15	35	40	10	40
Noticias y política	60	45	40	60	40	50
Ocio	25	20	15	5	20	10
Viajes y eventos	20	15	25	5	30	10
Macro-Averaging	**40,38**	**28,46**	**39,62**	**40,38**	**31,92**	**38,85**
Micro-Averaging	**40,38**	**28,46**	**39,62**	**40,38**	**31,92**	**38,85**

Tabla 36 - Cobertura en categorías - Resultados Finales

Mientras que la precisión depende del número de resultados dados, la cobertura es más dependiente del tipo de documento, se aprecia, que aquellos documentos que son completos o tienen las 100 palabras más repetidas obtienen una cobertura máxima mientras que aquellos que únicamente tienen los 30 primeros segundos no alcanzan dicha cobertura. Esto es normal, ya que existirán documentos tan pequeños que el clasificador no pueda obtener datos suficientes para su clasificación. Por lo tanto, el tamaño y tipo de documento influye en el clasificador.

Las dos tablas anteriores (Tabla 35 y Tabla 36) muestran un resumen del comportamiento de los distintos clasificadores en cuanto a las distintas categorías. Se observa claramente que parece observarse una clara relación entre las mismas categorías y los distintos tipos de sistemas, es decir, aquellas categorías que obtienen en un tipo de sistema una precisión o cobertura alta, también lo obtienen en el resto de sistemas. Mientras la precisión es mayor por lo general en los sistemas basados en ejemplos la cobertura es mayor en los sistemas basados en patrones.

Para terminar con las conclusiones sobre las pruebas, es necesario hablar sobre las categorías y cómo el sistema se confunde entre ellas. Las distintas matrices de confusión que se han estudiado permiten hacer un breve resumen sobre este comportamiento. En concreto, analizando la Tabla 16, que muestra la matriz de confusión de la configuración escogida, donde se aprecian diferentes características del sistema.

Para resumir y a la vez refrescar los resultados obtenidos se muestra la siguiente tabla.

Categorías	Predicha		Real		M. Confusión (diagonal)	P1 (%)	R1 (%)	F1 (%)
	N	%	N	%				
Automoción	16	6,15	20	7,69	3	17,65	15	17,05
Ciencia	26	10	20	7,69	16	66,67	80	68,97
Cine	21	8,08	20	7,69	4	18,18	20	18,52
Comedia	25	9,62	20	7,69	6	22,22	30	23,44
Deportes	23	8,85	20	7,69	5	22,73	25	23,15
Educación	18	6,92	20	7,69	11	57,89	55	57,29
Gente y	12	4,62	20	7,69	13	30,77	20	27,78
Instrucciones	17	6,54	20	7,69	14	82,35	70	79,55
Juegos	**21**	**8,08**	**20**	**7,69**	**18**	**85,71**	**90**	**86,54**
Mascotas	10	3,85	20	7,69	4	45,45	25	39,06
Noticias	32	12,3	20	7,69	12	37,93	55	40,44
Ocio	21	8,08	20	7,69	5	15,79	15	15,63
Viajes	18	6,92	20	7,69	4	21,05	20	20,83

Tabla 37 - Resumen de las características por categoría.

Esta tabla muestra, en porcentaje y número, la cantidad de documentos que son de una categoría (real) frente a la cantidad de documentos que el sistema ha clasificado en una categoría (predicho) y frente a su correspondiente R1, P1 y F1. También muestra la diagonal principal de la matriz de confusión, es decir, el número de aciertos (TP) de cada categoría.

Por ejemplo, la categoría Noticias es una categoría conflictiva, el sistema confunde 20 (32-12) vídeos de las otras categorías con ésta. En contrapunto está la categoría Juegos que, confunde únicamente 3 vídeos de los 21 que asigna a la misma por lo que tiene mucha precisión.

Algo importante de analizar es la estrecha relación entre el tipo de categorías y las que tienen precisiones altas, es decir, aquellas categorías específicas, tipo Juegos, Ciencia, Educación, Instrucciones, que están más definidas en un campo semántico concreto, obtienen precisiones y coberturas altas. Mientras que otras que no están tan definidas (Viajes, Gente y Noticias, por ejemplo) poseen una precisión baja.

Por ejemplo, observando la categoría Viajes se observa (Tabla 16) que la confunde demasiadas veces (5) con Ocio. Estos errores se dan porque gran parte de los vídeos de Ocio son en realidad sobre viajes y otros de este tipo.

Analizando las anteriores tablas y observando los distintos diagramas de barras, se puede concluir que la mejor configuración es aquella que está basada en patrones, los documentos tienen las 100 palabras más repetidas y toma como parámetro el valorar 50 palabras de cada documento.

CAPÍTULO 5 - CONCLUSIONES Y TRABAJOS FUTUROS

5.1 CONCLUSIONES

La clasificación automática de vídeos es un campo de reciente interés y por tanto poco estudiado hasta el momento. Los distintos sistemas implementados ofrecen eficiencias y resultados bastantes bajos. La actual sociedad hace que cada vez los documentos audiovisuales sean más importantes, esto unido a la creciente demanda de contenidos de este tipo en internet, están haciendo que aumente el número de trabajos de investigación en este campo.

La clasificación de vídeos, ya sea por contenido como por temática es un problema a afrontar y el presente documento ha realizado un estudio sobre este tema. Este tipo de búsquedas y clasificaciones están en proceso de investigación y por lo tanto hoy día ofrecen resultados bastante ineficientes.

No obstante, el presente estudio ha permitido concluir ciertos puntos a la vez que analizar ciertas características de estos sistemas y de lo que queda por investigar.

Las actuales técnicas de clasificación automática de documentos permiten clasificar textos como noticias de prensa y vídeos de otro tipo, con precisiones superiores al 90%. La clave de estos sistemas es la baja tasa de errores debido a unos corpus de entrenamiento realizados por expertos donde no existen fallos de conceptos ni semánticos, permitiendo así al clasificador realizar su labor matemática casi perfectamente. Una base de datos de vídeos más completa o mejor clasificada, por expertos, podría haber ayudado a los distintos sistemas evaluados en el actual estudio.

Los sistemas actuales de clasificación basada en texto, aparte del algoritmo de aprendizaje, se apoyan además en distintos métodos basados en reglas, así como otros, para obtener mejores precisiones y eliminar posibles errores. El actual estudio ha demostrado que el factor semántico, la ambigüedad de las distintas categorías, son problemas más que importantes, ya que, es incluso difícil de manera manual, clasificar algunos vídeos en una u otra categoría.

Los sistemas de transcripción automática que existen en la actualidad son sistemas eficientes pero suelen estar basados en el hablante y no suelen ser eficientes en conversaciones, en ambientes con música, ruidos y otros como audios de baja calidad. Los

errores cometidos en la fase de transcripción automática hacen que el clasificador los arrastre hasta el final.

El actual proyecto ha conseguido unos resultados aceptables abriendo además varias líneas con las que poder jugar e investigar en futuros proyectos y conseguir mejores resultados. Una precisión y una cobertura superiores al 40% permiten afirmar que el proyecto ha conseguido una eficiencia aceptable, comparando estos resultados con los ofrecidos por Villena y Lana en sus estudios para CLEF, ya que su sistema quedó en segunda posición, y teniendo en cuenta que el sistema aquí desarrollado es para español y no inglés, se incrementa la dificultad dado que la conversión voz-texto es de peor calidad (en general).

5.2 TRABAJOS FUTUROS

Para terminar queda por decir lo que en este proyecto se puede mejorar y lo que puede ser una posible línea de actuación para mejorarlo. Por lo tanto, las áreas en las que se podría investigar y trabajar en un futuro serían:

- **Reconocimiento automático del habla:** La investigación de futuros trabajos en esta línea puede ayudar al sistema en varios puntos como la precisión y la velocidad.

  A medida que los sistemas de reconocimiento automático del habla mejoren, el sistema aumentará en su capacidad para clasificar. Actualmente estos sistemas poseen demasiadas limitaciones para poder afirmar que son aptos y no ofrecen una muy buena calidad para un sistema de clasificación automática de vídeos.

  Un sistema de reconocimiento automático del habla utilizado en un clasificador de vídeos no debe estar orientado al hablante y son precisamente estos sistemas los que peor rendimiento dan.
- **Preprocesado de los documentos:** El procesado que se aplica a los documentos una vez se tienen en formato texto es una de las áreas y campos que actualmente están más presentes en la investigación de la clasificación automática.

  La selección de términos importantes, la expansión de los mismos y su lematización puede ser un buen trabajo futuro a realizar. Técnicas como la indexación semántica latente podrían ayudar a mejorar los resultados del sistema, como se ha estudiado en otros trabajos.
- **Mejora del corpus de entrenamiento:** Un trabajo futuro que sería interesante realizar es la mejora de estos sistemas para obtener un corpus de entrenamiento mejor elaborado y con menos fallos.
- **Algoritmos de aprendizaje y clasificación automática**: Hoy día existen diversos algoritmos de clasificación automática de textos. El presente estudio ha utilizado un sistema (Lucene) que implementa un K-NN mejorado porque en un trabajo anterior se demostró que daba mejores resultados que el resto.

  No obstante, existen otros algoritmos comentados en este mismo trabajo que puede ser interesante implementar y analizar el comportamiento del clasificador.

- **Creación de un servicio de clasificación automática de vídeos:** El presente proyecto ha estudiado e implementado un sistema de clasificación de vídeos más bien con el objetivo de estudiar la viabilidad de este tipo de sistemas pero no ha estudiado una arquitectura capaz de clasificar grandes colecciones de documentos audiovisuales como servicio.

  Se pueden crear dos líneas para trabajos futuros que están estrechamente relacionadas.

  Existe la posibilidad de estudiar un sistema que ofrezca sugerencias de categorías o que clasifique automáticamente un vídeo en tiempo de ejecución. Un ejemplo sería integrar este módulo en un sitio web como YouTube, donde los usuarios simplemente subirían los vídeos y el sistema se encargaría de o bien sugerirles categorías o bien clasificarlo directamente.

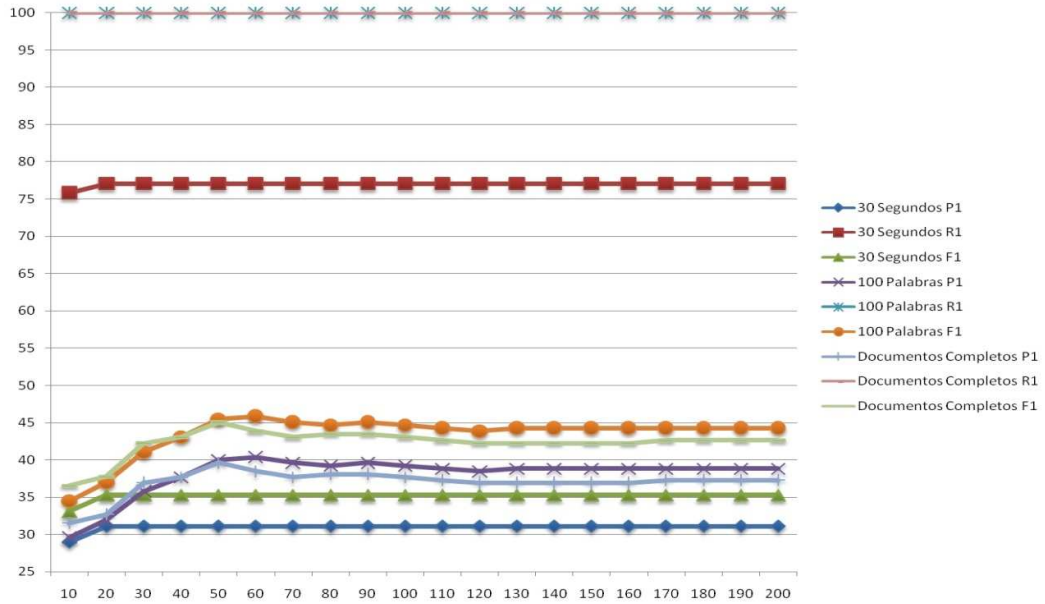
  Otra posibilidad sería el estudio de una arquitectura capaz de clasificar cientos de vídeos automáticamente sin la ayuda del ser humano. La aplicación directa de este sistema sería aplicable, por ejemplo, a todas las cadenas actuales de televisión o productoras con material audiovisual.
- **Clasificación automática no basada en texto:** Aunque es una línea de investigación todavía demasiado poco estudiada, una posibilidad es realizar este mismo proyecto estudiando las actuales técnicas de clasificación de imágenes, para poder así realizar un clasificador automático de vídeos basado en fotogramas, bien exclusivamente o bien como complemento al audio.

ANEXOS

Anexo 1 - Clasificación Basada en Patrones - Datos globales

Palabras	30 Segundos			100 Palabras			Documentos Completos		
	P1	R1	F1	P1	R1	F1	P1	R1	F1
10	28,99	75,82	33,08	29,62	100	34,47	31,54	100	36,54
20	31,09	77,08	35,31	31,92	100	36,95	32,69	100	37,78
30	31,09	77,08	35,31	35,77	100	41,04	36,92	100	42,25
40	31,09	77,08	35,31	37,69	100	43,06	37,69	100	43,06
50	31,09	77,08	35,31	40	100	45,45	39,62	100	45,06
60	31,09	77,08	35,31	**40,38**	**100**	**45,85**	38,46	100	43,86
70	31,09	77,08	35,31	39,62	100	45,06	37,69	100	43,06
80	31,09	77,08	35,31	39,23	100	44,66	38,08	100	43,46
90	31,09	77,08	35,31	39,62	100	45,06	38,08	100	43,46
100	31,09	77,08	35,31	39,23	100	44,66	37,69	100	43,06
110	31,09	77,08	35,31	38,85	100	44,26	37,31	100	42,66
120	31,09	77,08	35,31	38,46	100	43,86	36,92	100	42,25
130	31,09	77,08	35,31	38,85	100	44,26	36,92	100	42,25
140	31,09	77,08	35,31	38,85	100	44,26	36,92	100	42,25
150	31,09	77,08	35,31	38,85	100	44,26	36,92	100	42,25
160	31,09	77,08	35,31	38,85	100	44,26	36,92	100	42,25
170	31,09	77,08	35,31	38,85	100	44,26	37,31	100	42,66
180	31,09	77,08	35,31	38,85	100	44,26	37,31	100	42,66
190	31,09	77,08	35,31	38,85	100	44,26	37,31	100	42,66
200	31,09	77,08	35,31	38,85	100	44,26	37,31	100	42,66

Anexo 2 - Clasificación basada en patrones - Datos globales (gráfica)



Anexo 3 - Precisión (P1) Global - Clasificación Basada en Ejemplos - 100 Palabras más repetidas

P1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	35,4	35,8	35,8	35,4	35,4	35,8	35,4	35,4	35,4
20	36,2	36,9	36,9	36,9	36,5	36,9	36,5	36,5	36,5
30	34,6	35	35	35,4	35,4	35,4	35,4	35	35
40	38,5	38,8	38,5	38,5	38,8	38,5	38,8	38,8	38,5
50	40	40,4	40	40	40	40	40	39,6	40
60	38,1	38,5	38,1	38,5	38,1	38,5	38,5	37,7	37,7
70	34,6	35,4	35,4	35	35,4	35,4	35,4	35,4	35,4
80	36,2	35,8	36,2	36,2	36,2	35,8	35,8	35,8	36,5
90	37,7	37,7	36,9	36,9	36,9	36,9	37,7	37,7	37,7
100	35,8	35,4	35,4	35,4	35,4	35,4	35,8	35,8	35,8
110	35,4	35,4	35,4	35,4	35,4	35,4	35,8	35,8	35,8
120	35,8	35,8	36,2	36,2	36,2	36,2	36,5	36,5	36,5
130	35,8	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2
140	36,2	36,2	35,8	35,4	35,4	35,4	35,8	35,8	35,8
150	36,2	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2
160	35,8	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2
170	35,8	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,5
180	35,8	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2
190	35,4	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2
200	35,4	35,8	35,8	35,8	35,8	35,8	36,2	36,2	36,2

Anexo 4 - Cobertura (R1) Global - Clasificación Basada en Ejemplos - 100 Palabras más repetidas

R1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	100	100	100	100	100	100	100	100	100
20	100	100	100	100	100	100	100	100	100
30	100	100	100	100	100	100	100	100	100
40	100	100	100	100	100	100	100	100	100
50	100	100	100	100	100	100	100	100	100
60	100	100	100	100	100	100	100	100	100
70	100	100	100	100	100	100	100	100	100
80	100	100	100	100	100	100	100	100	100
90	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
110	100	100	100	100	100	100	100	100	100
120	100	100	100	100	100	100	100	100	100
130	100	100	100	100	100	100	100	100	100
140	100	100	100	100	100	100	100	100	100
150	100	100	100	100	100	100	100	100	100
160	100	100	100	100	100	100	100	100	100
170	100	100	100	100	100	100	100	100	100
180	100	100	100	100	100	100	100	100	100
190	100	100	100	100	100	100	100	100	100
200	100	100	100	100	100	100	100	100	100

Anexo 5 - Medida-F (F1) Global - Clasificación Basada en Ejemplos - 100 Palabras más repetidas

F1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	40,6	41	41	40,6	40,6	41	40,6	40,6	40,6
20	41,4	42,3	42,3	42,3	41,9	42,3	41,9	41,9	41,9
30	39,8	40,2	40,2	40,6	40,6	40,6	40,6	40,2	40,2
40	43,9	44,3	43,9	43,9	44,3	43,9	44,3	44,3	43,9
50	45,5	45,9	45,5	45,5	45,5	45,5	45,5	45,1	45,5
60	43,5	43,9	43,5	43,9	43,5	43,9	43,9	43,1	43,1
70	39,8	40,6	40,6	40,2	40,6	40,6	40,6	40,6	40,6
80	41,4	41	41,4	41,4	41,4	41	41	41	41,9
90	43,1	43,1	42,3	42,3	42,3	42,3	43,1	43,1	43,1
100	41	40,6	40,6	40,6	40,6	40,6	41	41	41
110	40,6	40,6	40,6	40,6	40,6	40,6	41	41	41
120	41	41	41,4	41,4	41,4	41,4	41,9	41,9	41,9
130	41	41	41	41	41	41	41,4	41,4	41,4
140	41,4	41,4	41	40,6	40,6	40,6	41	41	41
150	41,4	41	41	41	41	41	41,4	41,4	41,4
160	41	41	41	41	41	41	41,4	41,4	41,4
170	41	41	41	41	41	41	41,4	41,4	41,9
180	41	41	41	41	41	41	41,4	41,4	41,4
190	40,6	41	41	41	41	41	41,4	41,4	41,4
200	40,6	41	41	41	41	41	41,4	41,4	41,4

Anexo 6 - Precisión (P1) Global - Clasificación Basada en Ejemplos - 30 Primeros segundos

P1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	29,8	30,3	30,3	29,8	29,8	29,8	29,8	29,8	29,8
20	33,2	33,2	34	34,5	34,5	34,5	34,5	34,5	34,5
30	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
40	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
50	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
60	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
70	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
80	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
90	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
100	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
110	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
120	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
130	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
140	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
150	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
160	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
170	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
180	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
190	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9
200	33,6	34	34,5	34,9	34,9	34,9	34,9	34,9	34,9

Anexo 7 - Cobertura (R1) Global - Clasificación Basada en Ejemplos - 30 Primeros segundos

R1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	76,3	76,6	76,6	76,3	76,3	76,3	76,3	76,3	76,3
20	78,2	78,2	78,6	78,8	78,8	78,8	78,8	78,8	78,8
30	78,4	78,6	78,8	79	79	79	79	79	79
40	78,4	78,6	78,8	79	79	79	79	79	79
50	78,4	78,6	78,8	79	79	79	79	79	79
60	78,4	78,6	78,8	79	79	79	79	79	79
70	78,4	78,6	78,8	79	79	79	79	79	79
80	78,4	78,6	78,8	79	79	79	79	79	79
90	78,4	78,6	78,8	79	79	79	79	79	79
100	78,4	78,6	78,8	79	79	79	79	79	79
110	78,4	78,6	78,8	79	79	79	79	79	79
120	78,4	78,6	78,8	79	79	79	79	79	79
130	78,4	78,6	78,8	79	79	79	79	79	79
140	78,4	78,6	78,8	79	79	79	79	79	79
150	78,4	78,6	78,8	79	79	79	79	79	79
160	78,4	78,6	78,8	79	79	79	79	79	79
170	78,4	78,6	78,8	79	79	79	79	79	79
180	78,4	78,6	78,8	79	79	79	79	79	79
190	78,4	78,6	78,8	79	79	79	79	79	79
200	78,4	78,6	78,8	79	79	79	79	79	79

Anexo 8 - Medida-F (F1) Global - Clasificación Basada en Ejemplos - 30 Primeros segundos

F1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	34	34,4	34,4	34	34	34	34	34	34
20	37,5	37,5	38,4	38,8	38,8	38,8	38,8	38,8	38,8
30	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
40	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
50	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
60	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
70	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
80	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
90	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
100	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
110	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
120	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
130	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
140	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
150	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
160	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
170	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
180	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
190	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3
200	38	38,4	38,8	39,3	39,3	39,3	39,3	39,3	39,3

Anexo 9 -Precisión (P1) Global - Clasificación Basada en Ejemplos - Documentos completos

P1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	33,8	35,4	35	35	35	35	35	35	35
20	35,8	35,8	35,4	35,8	36,2	35,8	35,8	36,2	36,2
30	35,4	35,8	36,2	36,2	36,2	36,2	36,2	36,2	36,2
40	38,8	38,1	38,5	38,5	38,5	38,5	38,5	38,8	38,8
50	37,7	37,7	37,3	37,3	37,3	37,7	38,1	38,1	38,1
60	36,5	37,3	36,9	36,5	36,9	37,3	37,3	37,3	37,3
70	34,6	34,2	35	35	34,6	35	34,6	34,6	34,6
80	36,2	36,5	36,5	36,5	36,2	36,2	36,2	36,2	36,2
90	35,4	35,8	36,2	36,2	36,2	36,2	36,2	35,8	35,8
100	35,4	35,4	35,8	35,8	35,8	36,2	36,2	35,8	36,2
110	34,2	34,6	35	35	35	35	35	34,6	34,6
120	35	35	35,4	35,4	35,4	35,4	35,4	35	35
130	35	35,4	35,8	35,8	35,8	35,8	35,8	35,4	35,4
140	35,4	35,8	35,8	35,8	35,8	35,8	35,8	35,4	35,4
150	35	35	35,4	35,4	35,8	35,8	35,8	35,4	35,4
160	34,6	34,6	35	35,4	35,4	35,4	35,4	35	35
170	35	35	35,8	35,8	35,8	35,8	35,8	35,4	35,4
180	34,6	35	35,4	35,4	35,4	35,4	35,8	35,4	35,4
190	34,2	35	35,4	35,4	35,4	35,4	35,4	35	35
200	34,2	34,2	35	35	35	35	35	34,6	34,6

Anexo 10 - Cobertura (R1) Global - Clasificación Basada en Ejemplos - Documentos completos

R1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	100	100	100	100	100	100	100	100	100
20	100	100	100	100	100	100	100	100	100
30	100	100	100	100	100	100	100	100	100
40	100	100	100	100	100	100	100	100	100
50	100	100	100	100	100	100	100	100	100
60	100	100	100	100	100	100	100	100	100
70	100	100	100	100	100	100	100	100	100
80	100	100	100	100	100	100	100	100	100
90	100	100	100	100	100	100	100	100	100
100	100	100	100	100	100	100	100	100	100
110	100	100	100	100	100	100	100	100	100
120	100	100	100	100	100	100	100	100	100
130	100	100	100	100	100	100	100	100	100
140	100	100	100	100	100	100	100	100	100
150	100	100	100	100	100	100	100	100	100
160	100	100	100	100	100	100	100	100	100
170	100	100	100	100	100	100	100	100	100
180	100	100	100	100	100	100	100	100	100
190	100	100	100	100	100	100	100	100	100
200	100	100	100	100	100	100	100	100	100

Anexo 11 - Medida-F (F1) Global - Clasificación Basada en Ejemplos - Documentos completos

F1									
Palabras	10 Vídeos	15 Vídeos	20 Vídeos	25 Vídeos	30 Vídeos	35 Vídeos	40 Vídeos	45 Vídeos	50 Vídeos
10	39	40,6	40,2	40,2	40,2	40,2	40,2	40,2	40,2
20	41	41	40,6	41	41,4	41	41	41,4	41,4
30	40,6	41	41,4	41,4	41,4	41,4	41,4	41,4	41,4
40	44,3	43,5	43,9	43,9	43,9	43,9	43,9	44,3	44,3
50	43,1	43,1	42,7	42,7	42,7	43,1	43,5	43,5	43,5
60	41,9	42,7	42,3	41,9	42,3	42,7	42,7	42,7	42,7
70	39,8	39,4	40,2	40,2	39,8	40,2	39,8	39,8	39,8
80	41,4	41,9	41,9	41,9	41,4	41,4	41,4	41,4	41,4
90	40,6	41	41,4	41,4	41,4	41,4	41,4	41	41
100	40,6	40,6	41	41	41	41,4	41,4	41	41,4
110	39,4	39,8	40,2	40,2	40,2	40,2	40,2	39,8	39,8
120	40,2	40,2	40,6	40,6	40,6	40,6	40,6	40,2	40,2
130	40,2	40,6	41	41	41	41	41	40,6	40,6
140	40,6	41	41	41	41	41	41	40,6	40,6
150	40,2	40,2	40,6	40,6	41	41	41	40,6	40,6
160	39,8	39,8	40,2	40,6	40,6	40,6	40,6	40,2	40,2
170	40,2	40,2	41	41	41	41	41	40,6	40,6
180	39,8	40,2	40,6	40,6	40,6	40,6	41	40,6	40,6
190	39,4	40,2	40,6	40,6	40,6	40,6	40,6	40,2	40,2
200	39,4	39,4	40,2	40,2	40,2	40,2	40,2	39,8	39,8

REFERENCIAS

Las referencias incluidas en este apartado a continuación están ordenadas alfabéticamente.

Apache, 2009

Apache, http://www.apache.org/ (Última visita 29/11/2009).

Bacan, Pandzic, & Gulija, 2005

Hrvoje Bacan, Igor S. Pandzic y Darko Gulija. Automated News Item Categorization. http://www.ii.ist.i.kyoto-u.ac.jp/jsai2005ws/proceedings/bacan.pdf (Última visita 28/10/2009).

Cabello Pardos, 2004.

Cabello Pardos, E. (2004). Técnicas de Reconocimiento Facial mediante Redes Neuronales. (Tesis Doctoral). http://oa.upm.es/215/1/10200404.pdf (Última visita 28/10/2009).

Calvo, 2000

R.H. Calvo, H.A. Ceccatto. Clasificación inteligente de documentos. http://www.ee.usyd.edu.au/~rafa/papers/ToDo/jaio2k/icie00r.rtf (Última visita 22/09/2009).

Chan, M.L., 1981

Lois Mais Chan, Cataloging and classification: an introduction. McGraw-Hill.

CLEF, 2009

Cross Language Evaluation Forum. http://clef-campaign.org/ (Última visita 22/09/2009

Collada Pérez, 2009

Sonia Collada Pérez, Sistema de indexación y búsqueda de documentos audiovisuales. Universidad Carlos III. Departamento de Ingeniería Telemática. Proyecto de Fin de Carrera.

Cortijo Bon, 2000

Cortijo Bon, F. J. (2000). Técnicas supervisadas II: Aproximación no paramétrica. http://iie.fing.edu.uy/ense/asign/recpat/material/tema3_00-01/ (Última visita 21/09/09).

Dagan, Glickman, Magnini, 2005.

Dagan, I., Glickman, O., & Magnini, B. (2005). The PASCAL Recognising Textual Entailment. Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment. http://u.cs.biu.ac.il/~nlp/RTE1/Proceedings/dagan_et_al.pdf (Última visita 28/10/2009).

David & Lerner, 2004

David, A., & Lerner, B. (2004). Support Vector Machine-Based Image Classification for Genetic Syndrome Diagnosis. Elsevier Science Inc. http://portal.acm.org/citation.cfm?id=1088404 (Última visita 28/10/2009).

Dragon, 2009

Dragon NaturallySpeaking. http://www.nuance.com/naturallyspeaking/ (Última visita 28/11/2009).

Férnández, 2006

Jesús Fernández Benito, Sistema de Question de Question Answering Basado en Wikipedia. Universidad Carlos III de Madrid.

Figuerola y otros, 2004

Figuerola, C. G., Alonso Berrocal, J. L., Zazo Rodriguez, A. F., & Rodriguez, E. (2004). Algunas Técnicas de Clasificación Automática de Documentos. http://multidoc.rediris.es/cdm/viewarticle.php?id=28&layout=html (Última visita 20/09/2009).

Gallo, San-Segundo, 2009

B. Gallo, R. San-Segundo, J.M. Lucas, R. Barra, L.F. D'Haro, F. Fernández. Aplicación de métodos estadísticos para la traducción de voz a Lengua de Signos. http://rua.ua.es/dspace/bitstream/10045/8603/1/PLN_41_30.pdf (Última visita 28/10/2009).

Gavaldà i Camps, 2009

Marsal Gavaldà i Camps. La investigación en tecnologías de la lengua. http://www.prbb.org/quark/19/019021.htm (Última visita 20/05/2009).

Google, 2009

Google TM, http://www.google.es/corporate . (Última visita 16/09/2009).

Ide, 1999

Ichiro Ide, Koji Yamamoto y Hidehiko Tanaka. Automatic Video Indexing based on Shot Classification. http://biblioteca.universia.net/ficha.do?id=41563267 (Última visita 16/09/2009).

Java, 2009

Java, http://www.java.com/ (Última visita 10/09/2009).

Jeong, 2002

C.Y. Jeong, S.W. Han y T.Y. Nam. Automatic Objectionable Video Classification System. http://portal.acm.org/citation.cfm?id=1369648 (Última visita 10/06/2009).

JULIUS, 2009

JULIUS 4.2. http://julius.sourceforge.jp/ (Última visita 10/11/2009).

LEMUR, 2009

LEMUR 4.10. http://www.lemurproject.org/ (Última visita 22/11/2009).

Llisterri, 2008

Joaquin Llisterri, Universidad autónoma de Barcelona, Las tecnologías lingüísticas. http://liceu.uab.es/~joaquim/language_technology/HLT/tecnol_ling_gen.html (Última visita 22/10/2009).

López Herrera, 2005

Antonio Gabriel López Herrera. Modelos de Sistemas de Recuperación de Información Documental Basados en Información Lingüística Difusa. Tesis Doctoral. Universidad de Granada (2005). http://hera.ugr.es/tesisugr/15894046.pdf (Última visita 28/10/2009).

Lu, 2002

Chen Lu, Mark S. Drew y James Au. An Automatic Video Classification System Based on a Combination of HMM and Video. http://www.cs.sfu.ca/~mark/ftp/IntJSmartEnggSysDesign02/intjsmenggsysdes02.pdf (Última visita 23/11/2009).

Lucene, 2009

Apache Lucene 3.0.0. http://lucene.apache.org/java/docs/ (Última visita 30/11/2009).

Luque Rodríguez, 2006

María Luque Rodríguez. Modelos de Recuperación de la Información basados en Información Lingüística Difusa y Algoritmos Evolutivos. Mejorando la Representación de las Necesidades de Información. Tesis Doctoral. Universidad de Granada (2006). http://hera.ugr.es/tesisugr/15350605.pdf (Última visita 28/10/2009).

Media Mining Indexer, 2009

Media Mining Indexer. Sail Labs Tecnology. http://www.sail-technology.com/products/commercial-products/media-mining-indexer.html (Última visita 28/10/2009).

Molina Félix, 2002

Luis Carlos Molina Félix, *Data mining*: torturando a los datos hasta que confiesen, http://www.uoc.edu/web/esp/art/uoc/molina1102/molina1102.html (Última visita 16/09/2009)

Moreno, 2009

Asunción Moreno, La lengua española y las nuevas tecnologías. Inteligencia artificial y lengua española, http://cvc.cervantes.es/obref/congresos/sevilla/tecnologias/mesaredon_moreno.htm (Última visita 21/09/2009).

MPlayer, 2009

MPlayer. http://www.mplayerhq.hu/design7/news.html (Última visita 01/05/2009).

Nadeu, 2001

Climent Nadeu. Representación de la voz en el reconocimiento automático del habla, http://www.prbb.org/quark/21/021063.htm (Última visita 22/11/2009).

Nogueiras, 1999

Albino Nogueiras Rodríguez, 1999. Entrenamiento discriminativo de modelos ocultos de Markov de unidad subléxica para su aplicación a sistemas de reconocimiento automático del habla continua. Universidad Politécnica de

Cataluña. http://www.tesisenxarxa.net/TESIS_UPC/AVAILABLE/TDX-0722109-100044//TANR.pdf (Última visita 20/11/2009).

Novoa , Ballen, 2007

Novoa, D., & Ballen, L. (2007) La Indexación Semántica Latente en la Recuperación de Información. http://eprints.rclis.org/9867/1/ISL_Indizacion_semantica_Latente.pdf (Última visita 28/10/2009).

Olaso, 2002

Javier Mikel Olaso Fernández. Algoritmos de búsqueda en el reconocimiento automatic del habla. http://gtts.ehu.es/dEyE/Actualizable/Anual/Curso06-07/VII_Jornadas_IE/trabajos_dirigidos/I-JavierOlaso.pdf (Última visita 11/10/2009).

Papineni, Roukos, 2002

Papineni K., S. Roukos, T.Wardm W.J.Zhu. "BLEU: a method for automatic evaluation of machine translation". 40th Annual Meeting of the ACL, Philadelphia, PA, pp. 311-318. http://www.aclweb.org/anthology/P/P02/P02-1040.pdf (Última visita 28/10/2009).

Paz, 2007

Vadim Paz Madrid Gerelov et al. Librerías Lucene y dotLucene para Recuperación de Información. Estudio y desarrollo de casos prácticos. Departamento de Informática y Automática, Universidad de Salamanca, 2007. http://reina.usal.es/papers/pazmadrid2007librerias.pdf (Última visita 11/08/2009).

Perea-Ortega, 2008

José M. Perea-Ortega, Arturo Montejo Ráez, M. Teresa Martín Valdivia, Manuel C. Díaz Galiano, L. Alfonso Ureña López. SINAI at VideoCLEF 2008. http://www.clef-campaign.org/2008/working_notes/Perea-Ortega-paperVideoCLEF2008.pdf (Última visita 19/06/2009).

PHP, 2009

PHP 5.3.1 http://php.net/index.php (Última visita 03/12/2009)

Porter, 2008

Porter, Martin. Snowball stemmers and resources page. http://www.snowball.tartarus.org (Última visita 28/10/2009).

Resendiz, 2006

Juan Ángel Resendiz Trejo. Las máquinas de vectores de soporte para identificación en línea. http://www.ctrl.cinvestav.mx/~yuw/pdf/MaTesJAR.pdf (Última visita 08/04/2009).

RAE, 2009

Real Academia Española. http://www.rae.es/rae.html (Última visita 13/09/09).

Sánchez Jiménez, 2007

Rodrigo Sánchez Jiménez. La documentación en el proceso de evaluación de Sistemas de Clasificación automática, Departamento de biblioteconomía y documentación, Universidad Complutense de Madrid, http://revistas.ucm.es/inf/02104210/articulos/DCIN0707110025A.PDF (Última visita 28/10/2009).

Sebastiani, 2002

Fabrizio Sebastiani, Machine Learning in Automated Text Categorization. http://www.isti.cnr.it/People/F.Sebastiani/Publications/ACMCS02.pdf (Última visita 28/10/2009).

Sphinx, 2009

Sphinx 0.9. http://www.sphinxsearch.com/ (Última visita 28/11/2009).

Turner, K & Thost, 1995

Order statistics Combiners for neural classifiers. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.55.7361&rep=rep1&type=pdf (Última visita 28/10/2009).

Vapnik, 1995

Vapnik, V. N. (1995). The Nature of Statistical Learning Theory. Springer. http://cscs.umich.edu/~crshalizi/reviews/vapnik-nature/ (Última visita 28/10/2009).

Vazquez, Kozareva, Montoyo, 2006.

Vazquez, S., Kozareva, Z., & Montoyo, A. (2006). Contribución de la Información Semántica en un Sistema de Aprendizaje Automático para Resolver la Implicación

Textual. http://www.sepln.org/revistaSEPLN/revista/37/24.pdf (Última visita 28/10/2009).

Venegas, 2007

René Venegas, Clasificación de textos académicos en función de su contenido léxico-semántico. http://www.postgradolinguistica.ucv.cl/pr_curriculum_eve_doc.php?did=310 (Última visita 28/10/2009).

ViaVoice, 2009

IBM ViaVoice 10.0 http://www.nuance.com/viavoice/ (Última visita 30/10/2009).

VideoCLEF, 2008

Evaluation of Cross-Language Video Access. http://www.cdvp.dcu.ie/VideoCLEF/ (Última visita 23/09/2009).

Villena, Lana, 2008

Julio Villena Román y Sara Lana Serrano. MIRACLE at VideoCLEF 2008: Classification of Multilingual Speech Transcripts. http://www.daedalus.es/fileadmin/daedalus/doc/I%2BD/Villena2-paperCLEF2008_MIRACLE_Vid2RSS2008.pdf (Última visita 28/10/2009).

Villena Román, 2008.

Julio Villena Román. Inteligencia en Redes de Comunicaciones, Ingeniería de Telecomunicaciones, Universidad Carlos III de Madrid. 2008 http://www.it.uc3m.es/jvillena/irc/indice.html (Última visita 15/03/2009).

Wikipedia, 2009

Wikipedia, La enciclopedia Libre. http://es.wikipedia.org/wiki/Wikipedia:Acerca_de / (Última visita 01/12/2009).

YouTube, 2009

YouTube, http://www.YouTube.com/ (Última visita 01/12/2009).