GLOSARIO

Agente *software*

Un agente *software* es un programa autocontenido que opera como parte de un entorno, es capaz de tomar decisiones sobre la base de su percepción de dicho entorno y puede ejecutar acciones para alcanzar una serie de objetivos. Los agentes *software* actúan en representación de otros actores (agentes o, más comunmente, usuarios) y requieren nula o muy poca participación de los mismos para su funcionamiento.

Agrupamiento (*clustering*)

Véase **clasificación automática**.

Agrupamiento incremental/no incremental

Dos tipos de **clasificación automática**. El agrupamiento incremental trabaja sobre elementos aislados y parte del supuesto de que es posible considerarlos de uno en uno asignándolos a algún grupo ya disponible; por esta razón está especialmente indicada para conjuntos muy grandes. Por su parte, el agrupamiento no incremental opera sobre todo el conjunto de elementos, generalmente requiere comparar todos los elementos entre sí y, en consecuencia, es apta sólo para conjuntos relativamente pequeños.

Árbol de decisión

Una técnica de **categorización automática** en la que los nodos del árbol se corresponden con variables, los arcos con valores para las mismas y las hojas con las categorías predichas basándose en los valores de las variables que se encuentran en el recorrido entre la raíz del árbol y la hoja.

Automatic summarization

Véase **resumen automático**.

Autoridad

Según Kleinberg (1998) una autoridad es una página web fuertemente enlazada, o lo que es lo mismo, referenciada.

blindLight

Técnica estadística de **procesamiento de lenguaje natural** que establece métodos para la representación vectorial de textos escritos en cualquier idioma así como para la comparación de dichos vectores permitiendo el desarrollo de algoritmos de **clasificación**, **categorización**, **recuperación de información** y **resumen automático**. Estos vectores emplean *n*-gramas de caracteres a los que se asocia un valor de **significatividad** a partir tan sólo de los contenidos del documento original.

Boosting hypothesis

Hipótesis planteada por Kearns (1988) acerca de la posibilidad de construir un categorizador eficiente con una cadena de "categorizadores débiles" (aquellos cuya regla de decisión es sólo ligeramente mejor que una decisión tomada al azar). Se basa en el modelo de aprendizaje automático ***PAC***, según el cual el aprendizaje exitoso equivale a la minimización del error de la hipótesis (o regla de decisión).

Cadena léxica

Una cadena léxica es una secuencia de palabras semánticamente relacionadas que aparecen en un texto y que pueden ser adyacentes o encontrarse dispersas a lo largo del documento. Para encontrar dichas cadenas léxicas en un texto genérico es necesario utilizar recursos como *WordNet* que proporcionan la información necesaria sobre las posibles relaciones entre distintas palabras.

Categorización automática

Término que hace referencia a una amplia variedad de técnicas que tienen como objetivo asignar a un objeto dado una o más categorías (o etiquetas) de un conjunto predefinido. La categorización automática de documentos se realiza a partir del texto de los mismos y requiere una fase previa de entrenamiento durante la cual el categorizador es enfrentado con unos pocos ejemplos de las distintas categorías que debe reconocer.

Categorización mediante *boosting*

Una técnica de **categorización automática** basada en la denominada ***boosting hypothesis***. Se trata de un meta-algoritmo puesto que parte de la utilización de distintas técnicas de categorización. A diferencia de la **categorización mediante comités**, en el *boosting* los distintos categorizadores trabajan por etapas: (1) un categorizador se entrena sobre una parte del conjunto de entrenamiento y se prueba sobre el resto del conjunto; (2) aquellos documentos del conjunto de entrenamiento que clasifique mal, junto con algunos otros de su subconjunto de entrenamiento original, se utilizan para entrenar otro categorizador que, de este modo, "aprende" casos más "difíciles"; (3) este esquema se repite *n* veces.

Categorización mediante comités

Técnica de **categorización automática** basada en el uso simultáneo de varios categorizadores (un comité) que emiten un voto para cada elemento a categorizar. El resultado de dicha votación establece la categoría o categorías finalmente asignadas.

Categorizador bayesiano (*naïve Bayes*)

Técnica de **categorización automática** basada en el teorema de Bayes de probabilidad condicionada y que supone una total independencia entre las características que definen cada objeto. El apelativo de *naïve* se debe a lo irreal de esta suposición. La base de estos categorizadores es la siguiente: (1) a partir del conjunto de entrenamiento se puede establecer una probabilidad *a priori* para cada categoría y la probabilidad de cada característica condicionada para cada categoría; (2) para categorizar un objeto (definido por los valores de las características) basta con usar los datos anteriores para calcular la probabilidad de cada categoría condicionada a las características observadas en el objeto.

Centroide

Dado un conjunto de puntos multidimensionales (véase **modelo vectorial**) el centroide es aquel punto que tiene como coordenadas la media de los valores en cada dimensión.

Clasificación automática (agrupamiento o *clustering*)

Término que hace referencia a una amplia variedad de técnicas que, dentro de un conjunto de elementos, permiten identificar grupos que exhiben características similares (véase **similitud**). Para ello pueden dividir de manera iterativa el conjunto original en subconjuntos o comenzar por los elementos aislados e ir agrupando los más próximos. Así mismo, existe la posibilidad de operar sobre todo el conjunto simultaneamente o de manera paulativa (véase **Agrupamiento incremental/no incremental**).

Clustering (agrupamiento)

Véase **clasificación automática**.

Compresión de sentencias (*sentence compression*)

Técnica relacionada con el **resumen automático**. Según Knight y Marcu (2000) la compresión de sentencias tiene como finalidad conservar la información más relevante de una sentencia reescribiéndola en una forma más corta. El grado de sofisticación de estas técnicas es muy variable. Según Lin (2003) un **resumen extractivo** construido a partir de sentencias comprimidas no resulta necesariamente mejor que un resumen extractivo de la misma longitud sin compresión; aún así afirma que *"existe potencial en la compresión de sentencias pero es necesario encontrar un mejor sistema de compresión que tenga en cuenta para la optimización aspectos globales entre distintas sentencias"*.

Concentrador (*hub*)

Según Kleinberg (1998) un *hub* (o concentrador) es una página web que contiene enlaces a varias **autoridades**.

Consulta

Una consulta es la manifestación escrita de una necesidad de información formulada por un usuario a un sistema de **recuperación de información**.

Consulta informativa

Según Broder (2002) aquella **consulta** que un usuario formula en un buscador web para obtener algún tipo de información que supone está disponible en una o más páginas web (p.ej. `degenerative disc disease` o `muscle aches during pregnancy`).

Consulta navegacional

Según Broder (2002) aquella **consulta** que un usuario formula en un buscador web para llegar a un sitio web en particular (p.ej. `hotmail`, `renfe` o `universidad de oviedo`)

Consulta transaccional

Según Broder (2002) aquella consulta que un usuario formula en un buscador web para alcanzar un sitio web en el que llevar a cabo algún tipo de actividad (p.ej. `hotmail`, `weather` o `maps`)

Corpus

En lingüística un *corpus* es una colección, generalmente muy grande, de documentos (típicamente textos o habla transcrita) que muestran el uso real de una lengua natural. Un *corpus* puede contener muestras de un único lenguaje (*corpus* monolingüe) o de varios lenguajes (*corpus* multilingüe). En el caso de los *corpora* multilingües pueden ser comparables (el número de documentos y términos son similares para todos los idiomas y la temática es homogénea) o paralelos (los documentos son traducciones

literales y se dispone de información sobre su "alineación" a nivel de documento, párrafo o sentencia).

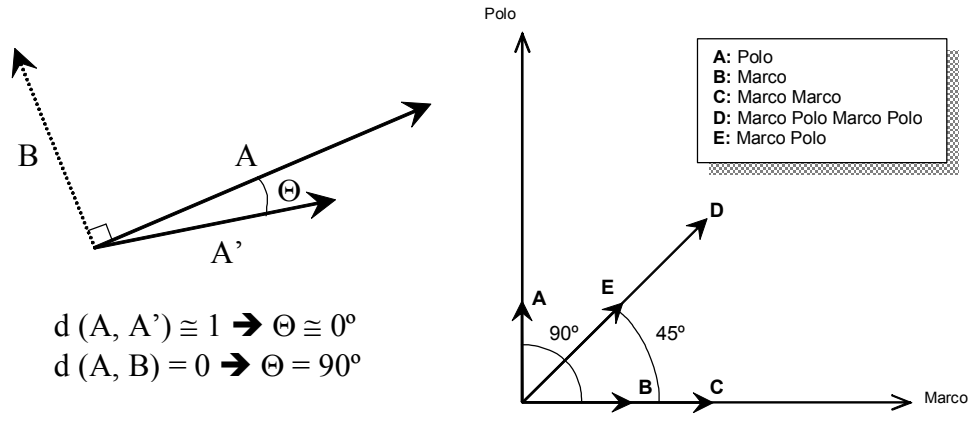
Coseno, función del (*cosine similarity*)

La función del coseno es una medida de **similitud** comunmente empleada en el **modelo vectorial**. Se calcula mediante la siguiente ecuación en la que n es el número de términos (dimensiones del espacio vectorial) y q_i y d_i son, respectivamente, el i-ésimo término de los documentos q y d.

$$\frac{\sum_{i=1}^{n} q_i \cdot d_i}{\sqrt{\sum_{i=1}^{n} q_i^2} \cdot \sqrt{\sum_{i=1}^{n} d_i^2}}$$

Puesto que en el modelo vectorial no se usan generalmente pesos negativos esta función ya es una similitud normalizada que, además, admite una interpretación geométrica sencilla: cuanto más próximo a 1 esté el valor obtenido más cercano a 0º será el ángulo formado por los vectores y, en consecuencia, más similares serán éstos; por el contrario, valores próximos a 0 implicará que los vectores son ortogonales (la máxima separación posible en un espacio vectorial en el que todos los términos toman valores positivos).

En la siguiente figura se muestran cinco documentos representados en un espacio vectorial de dos dimensiones así como los ángulos entre los vectores de algunas parejas ilustrativas. Obsérvese que para la función del coseno (y en general para cualquier medida de similitud) no se puede afirmar que una similitud total (en este caso un ángulo de 0º como el que forman *D* y E) implique la identidad entre ambos documentos. Compárense con los obtenidos con una medida de **disimilitud**.



Destilación de información

La destilación de información está relacionada en cierta medida con la **recuperación de información**; en el contexto de este trabajo hace referencia a técnicas como la respuesta de preguntas (*question answering*) o el **resumen automático** (*automatic summarization*).

Disimilitud

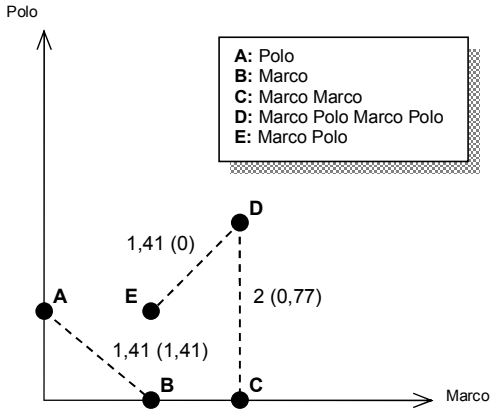
La disimilitud mide la discrepancia entre dos objetos a partir de sus características. Puesto que el **modelo vectorial** define un espacio multidimensional es posible determinar la disimilitud entre dos documentos i y j simplemente calculando la distancia entre ambos:

$$d_{ij} = \sqrt[p]{\sum_{k=1}^{n} \left(x_{ik} - x_{jk}\right)^p}$$

En la ecuación anterior, n es el número de términos o dimensiones del espacio, k es el término k-ésimo de cada documento y p es el orden de la distancia. Esta distancia se denominada de Minkowski y para p=1 equivale a la distancia Manhattan, para ese mismo valor pero datos binarios a la distancia de Hamming y para p=2 a la distancia euclídea.

No resulta demasiado adecuado emplear directamente la distancia puesto que se ve muy influida por el tamaño de los documentos. Esto puede paliarse en parte normalizando los vectores pero aun así no se dispondrá de una disimilitud normalizada (que varía entre 0 y 1). No obstante, a partir de las mismas características se puede calcular de manera directa una medida de **similitud** que, por otra parte, puede convertirse de manera trivial en una disimilitud normalizada si fuese necesario.

En la siguiente figura se muestran cinco documentos representados en un espacio vectorial de dos dimensiones así como las distancias euclídeas entre algunas parejas ilustrativas (entre paréntesis la distancia euclídea para los vectores normalizados). Obsérvese las diferencias entre los resultados obtenidos antes y después de la normalización de los vectores; compárense con los obtenidos con una medida de similitud como la **función del coseno**.



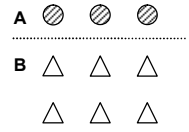
Entropía

La entropía es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) mediante la comparación con una clasificación previa (que proporciona una serie de clases). Dado un grupo S_r de tamaño n_r su entropía se define como:

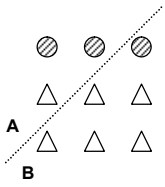
$$E(S_r) = -\sum_{i=1}^{q} p_{ir} \log p_{ir} = -\sum_{i=1}^{q} \frac{n_r^i}{n_r} \log \frac{n_r^i}{n_r}$$

donde q es el número de clases y n^i_r es el número de documentos de la clase i-ésima que fueron asignados al grupo r-ésimo. La entropía de la clasificación final se define como la suma de las entropías de todos los grupos ponderadas de acuerdo a su tamaño, es decir:

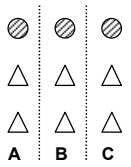
$$Entropia = \sum_{r=1}^{k} \frac{n_r}{n} E(S_r)$$



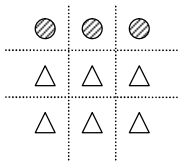
$$E(A) = -\left(\tfrac{3}{3} \cdot \log \tfrac{3}{3} + \tfrac{0}{3} \cdot \log \tfrac{0}{3}\right) = -(1 \cdot \log 1 + 0 \cdot \log 0) = 0$$
$$E(B) = -\left(\tfrac{0}{6} \cdot \log \tfrac{0}{6} + \tfrac{6}{6} \cdot \log \tfrac{6}{6}\right) = -(0 \cdot \log 0 + 1 \cdot \log 1) = 0$$
$$E = 0$$



$$E(A) = -\left(\tfrac{2}{3} \cdot \log \tfrac{2}{3} + \tfrac{1}{3} \cdot \log \tfrac{1}{3}\right) = -(-0{,}117 - 0{,}159) = 0{,}276$$
$$E(B) = -\left(\tfrac{1}{6} \cdot \log \tfrac{1}{6} + \tfrac{5}{6} \cdot \log \tfrac{5}{6}\right) = -(-0{,}130 - 0{,}066) = 0{,}196$$
$$E = \tfrac{3}{9} \cdot 0{,}276 + \tfrac{6}{9} \cdot 0{,}196 = 0{,}223$$



$$E(A) = E(B) = E(C) =$$
$$= -\left(\tfrac{1}{3} \cdot \log \tfrac{1}{3} + \tfrac{2}{3} \cdot \log \tfrac{2}{3}\right) = -(-0{,}159 - 0{,}117) = 0{,}276$$
$$E = 0{,}276$$



$$E(A) = E(B) = \ldots = E(H) = E(I) =$$
$$= -\left(\tfrac{1}{1} \cdot \log \tfrac{1}{1} + \tfrac{0}{1} \cdot \log \tfrac{0}{1}\right) = -(1 \cdot \log 0 + 0 \cdot \log 0) = 0$$
$$E = 0$$

Cuanto mayor es la semejanza de la solución obtenida y la clasificación externa menor es la entropía de dicha solución. El valor mínimo posible es 0 que supondría una clasificación idéntica a la externa o bien una solución trivial consistente en la división del conjunto de elementos en grupos formados por un único *ítem* (véase último caso en la figura).

Exhaustividad (*recall*)

La exhaustividad es una medida de la "calidad" de un sistema de **recuperación de información**. Se trata de la fracción del total de documentos relevantes que son obtenidos por un sistema *IR*. La exhaustividad máxima es, por definición, 1 y conlleva la recuperación de todos los documentos relevantes existentes. Esta medida por sí sola no es suficiente para caracterizar a un sistema *IR* puesto que se puede garantizar trivialmente una exhaustividad total recuperando todos los documentos de la colección para cualquier **consulta**. Por esa razón al evaluar un sistema siempre se calcula, además de la exhaustividad, su **precisión**. Ambas pueden combinarse en un solo valor: la **medida *F***. Véase además **curva de precisión-exhaustividad**.

Fallout (tasa o índice de irrelevancia, índice de fallos)

Una medida del rendimiento de un sistema de **recuperación de información** relacionada con la **precisión** y la **exhaustividad** aunque no tan utilizada como éstas. Se trata de la proporción de documentos no relevantes en la colección que se ofrecen como resultados de una consulta. Así pues, informa sobre la rapidez con que la precisión disminuye al aumentar la exhaustividad (o lo que es lo mismo, el número de documentos retornados por consulta).

Filtrado colaborativo

El filtrado colaborativo es una técnica que permite a un sistema sugerir a cada usuario en particular una selección de nuevos elementos sobre la base de sus preferencias en el pasado y de las valoraciones que, de dichos elementos, han hecho otros usuarios del sistema que coinciden, en mayor o menor medida, con las preferencias del usuario

original. Un ejemplo típico es el servicio de *Amazón* (`http://www.amazon.com`) *"Customers who bought this book also bought…"* ("Los clientes que compraron este libro también compraron…")

Folksonomía (*folksonomy*)

Término procedente del inglés *folksonomy* = *folk* + *taxonomy* que hace referencia a la categorización colaborativa de documentos (en general páginas web) mediante etiquetas (palabras clave) escogidas libremente por los usuarios. Un ejemplo del incipiente uso de folksonomías puede encontrarse en `http://del.icio.us`, un sitio web donde los usuarios almacenan sus enlaces favoritos etiquetándolos y pudiendo descubrir, a través de la exploración de dichas etiquetas, nuevos sitios web potencialmente relevantes para sus intereses.

Hub

Véase **concentrador**.

idf (*inverse document frequency*)

Método para ponderar los términos de un documento en un sistema de **recuperación de información**. Fue propuesto por Karen Spärck-Jones (1972) y se basa en la idea de que un término es tanto más informativo y, en consecuencia, importante cuanto menor es el número de documentos que lo contienen. Es decir, el peso de un término es inversamente proporcional al número de documentos que lo emplean. La expresión habitualmente empleada para el cálculo del valor *idf* es la siguiente:

$$w = -\log \frac{n}{N}$$

Donde w es el peso del término, n es el número de documentos que contienen dicho término y N es el total de documentos de la colección. Con frecuencia este método de ponderación se combina para formar el denominado ***tf*idf***.

Intersección Ω

En el contexto de la técnica ***blindLight*** se trata de un operador que permite la combinación de dos vectores documentales en un nuevo vector intersección. Dicho vector contiene los *n*-gramas que aparecen en los dos vectores originales y para cada uno de dichos *n*-gramas se asocia como significatividad el valor mínimo del par.

IR (*information retrieval*)

Véase **recuperación de información**.

Macropromediar y micropromediar (*macroaverage* vs. *microaverage*)

Dos formas de obtener resultados promedio al evaluar sistemas de **categorización automática** según Lewis (1991). Dado un conjunto de D documentos y una serie de K categorías un categorizador toma $D \cdot K$ decisiones que pueden ser evaluadas individualmente. A fin de ofrecer un único valor promedio puede obtenerse la precisión para cada categoría y posteriormente calcular su media o bien tomar todas las decisiones como un único conjunto. En el primer caso se habla de *macroaveraging* y en el segundo de *microaveraging*.

La diferencia entre una y otra medida es simple: en el caso de *microaveraging* tiene más influencia el resultado global (esto es, el número total de categorizaciones correctas) frente a las diferencias entre categorías (puede haber diferencias notables entre los resultados obtenidos para cada categoría) mientras que en el caso de *macroaveraging* influyen más las diferencias entre categorías que los resultados tomados en su

conjunto, es decir, se "premiaría" al categorizador que obtiene resultados similares en todas las categorías. En función del tipo de aplicación debe decidirse qué tipo de comportamiento es preferible y emplear un tipo u otro de promedio para la evaluación de los resultados.

A continuación se muestra un pequeño ejemplo. Supongamos una colección de documentos en distintos idiomas: 1000 en inglés (EN), 600 en español (ES), 300 en portugués (PT), 100 en alemán (DE) y 20 en francés (FR). Las categorías serían naturalmente los nombres de los idiomas y la lengua en que está escrito cada documento no se conoce *a priori*, es decir, deben categorizarse los documentos de acuerdo a su idioma. Supongamos que los resultados obtenidos al categorizar automáticamente dicha colección son los siguientes:

Categoría	Resultados		Precisión
EN	950 (EN)	50 (DE)	0,95
ES	600 (ES)	100 (PT)	0,86
PT	150 (PT)	10 (FR)	0,94
DE	50 (DE)	50 (EN)	0,50
FR	10 (FR)	50 (PT)	0,17
	1760	260	**Macro: 0,68**
	Micro: 0,87		

En este caso el valor macropromedio es de 0,68 y el micropromedio de 0,87. Es decir, un 87% del total de documentos fueron clasificados correctamente; sin embargo, puesto que el valor macropromediado es mucho menor puede afirmarse que este sistema se comporta de manera sustancialmente diferente frente a las distintas categorías.

Mapas Auto-organizativos (*Self-Organizing Maps* o *SOM*)

Véase ***Self-Organizing Maps***.

Máquinas de Vectores Soporte

Véase ***Support Vector Machines***.

Medida *F*

La medida *F* fue propuesta por van Rijsbergen (1979) como una medida única para evaluar la calidad de un sistema de **recuperación de información**. Esta medida combina en un solo valor la **precisión** y **exhaustividad** de un sistema:

$$F = \frac{1}{\alpha\left(\frac{1}{P}\right) + (1-\alpha)\left(\frac{1}{R}\right)}$$

En la ecuación anterior la precisión es *P* y la exhaustividad *R* (*recall*). Se suele utilizar un valor de α=½ por lo que la ecuación generalmente empleada para calcular la medida *F* es la siguiente:

$$F = 2\frac{P \cdot R}{P + R}$$

Medoide

Aquel elemento de un conjunto de puntos (véase **modelo vectorial**) que está más próximo a su **centroide**.

Modelo vectorial (*vector space model*)

Modelo propuesto por Salton y Lesk (1965) consistente en la representación de un conjunto de documentos como puntos en un entorno *T*-dimensional siendo *T* el número de términos distintos en el conjunto. Los términos son generalmente palabras o raíces o lemas de palabras. Cada documento será pues un vector de pesos, siendo estos nulos si el término no aparece en el documento y no nulos si el documento lo contiene; en este caso pueden usarse distintos métodos de ponderación, típicamente ***tf*idf***. Dada la naturaleza algebraica del modelo es posible definir **distancias** (y **similitudes**) entre los documentos siendo habitual el uso de la **función del coseno**.

Naïve Bayes

Véase **categorizador bayesiano**.

N-grama

Un *n*-grama es una secuencia de *n* elementos, palabras o caracteres, extraídos de un texto de forma no necesariamente correlativa. En el contexto de este trabajo se entiende por *n*-grama una secuencia de *n* caracteres contiguos que puede contener blancos y, por tanto, estar formado por segmentos de varias palabras consecutivas. Por ejemplo, los cinco primeros 3-gramas de esta definición serían `Un_`, `n_n`, `_n-`, `n-g` y `-gr` (se han reemplazado los blancos por guiones bajos).

NLP (*Natural Language Processing*)

Véase **procesamiento de lenguaje natural**.

Ontología

Una ontología, en un contexto informático, es según Gruber (1993) *"la especificación de una conceptualización. Esto es, una descripción de los conceptos y relaciones que pueden existir para un agente o una comunidad de agentes"* . La **Web Semántica** se basa en el uso intensivo de ontologías definidas como *"un documento o fichero que define formalmente las relaciones entre términos; una ontología típica para la Web consta de una taxonomía y de un conjunto de reglas de inferencia"* (Berners-Lee, Hendler y Lassila 2001, p.4).

PAC (*Probably Approximately Correct*)

Modelo matemático de aprendizaje automático (aplicable, por tanto, a la **categorización automática**) que establece la equivalencia entre aprendizaje exitoso y la minimización del error de una hipótesis (o regla de decisión) obtenida a partir de ejemplos de entrenamiento tomados al azar. Las hipótesis (o reglas) aprendidas son aproximadas pues fallan para una fracción de ejemplares establecida de manera arbitraria.

PageRank

PageRank (Page *et al.* 1998) hace referencia a un algoritmo para el cálculo del "prestigio" de una página web así como al valor calculado por dicho algoritmo. La técnica se basa en la idea de citación o referencia según la cual un documento muy citado (o lo que es lo mismo, enlazado) será más prestigioso que otro menos citado o no citado en absoluto (véase **autoridad**). Sin embargo, a diferencia de otros métodos, *PageRank* no considera por igual todos los enlaces recibidos por un documento sino en función del valor numérico (también *PageRank*) del documento del que parte el enlace. De este modo el "prestigio" o autoridad se propaga mediante los enlaces de unos documentos a otros: el *PageRank* de una página se divide por el número de enlaces de salida y se "transfiere" a los documentos enlazados. Así, documentos que reciben muchos enlaces aunque de poco valor serán muy relevantes y documentos que reciben pocos enlaces pero desde páginas con *PageRank* elevado serán igualmente

importantes. *PageRank* es uno de los métodos que emplea el buscador web *Google* (`http://www.google.com/`) para determinar la relevancia de los documentos que satisfacen una **consulta**.

Palabras vacías

Véase ***stop words***.

Pasaje

Según Salton *et al.* (1996) cada uno de los *"fragmentos de texto [en que puede dividirse un documento] que exhiben consistencia interna y que pueden distinguirse del resto de texto circundante"*.

Pi (Π)

En el contexto de la técnica ***blindLight*** se define Π como el cociente de la **significatividad total** del vector **intersección** de una **consulta** y un documento entre la significatividad total del vector consulta. Π revela en qué medida la consulta queda "satisfecha" por la intersección entre ésta y un documento resultante. Esta medida puede combinarse con **Ro (P)** para construir distintas medidas de **similitud** adaptadas a las necesidades de las diferentes aplicaciones.

Precisión

La precisión es una medida de la "calidad" de un sistema de **recuperación de información**. Se trata de la fracción de documentos obtenidos por un sistema *IR* que son relevantes. La precisión máxima es, por definición, 1 y supone que todos los documentos recuperados son relevantes. Esta medida por sí sola no es suficiente para caracterizar a un sistema *IR* puesto que si no se retorna ningún documento la precisión sería 1 ya que no hay ningún resultado irrelevante. Por ello al evaluar un sistema siempre se calcula, además de la precisión, su **exhaustividad**. Ambas pueden combinarse en un solo valor: la **medida *F***. Véase además **precisión en *k***, **precisión media**, **precisión interpolada** y **curva de precisión-exhaustividad**.

Precisión en *k*

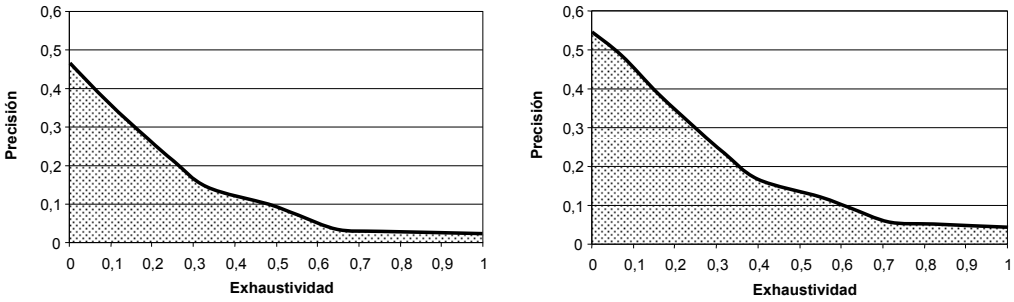
Se trata de la **precisión** de un sistema de **recuperación de información** cuando se han retornado exactamente *k* resultados. Por ejemplo, supongamos que un sistema *IR* retorna para una consulta 10 documentos de los cuales 4 son relevantes; entonces la precisión en 10 es de 0,4. Si se retornan 10 documentos más y ninguno es relevante entonces la precisión en 20 será de 0,2. Dicho de otro modo:

$$precisionEnK = \frac{\sum_{i=1}^{K} rel(i)}{K}$$

Donde *K* es el número de resultados, *i* es el resultado *i*-ésimo y *rel(i)* retorna 1 si el resultado *i*-ésimo es relevante para la consulta y 0 en caso contrario.

Precisión-Exhaustividad, curva de

Representación gráfica del comportamiento de un sistema de **recuperación de información** mostrando cómo varía la **precisión interpolada** (eje de ordenadas) con la **exhaustividad** (eje de abscisas). Una curva precisión-exhaustividad típica es cóncava y decreciente. Este tipo de representación permite comparar con relativa facilidad dos sistemas *IR* puesto que un mejor rendimiento (mayor precisión y exhaustividad) supone una mayor superficie encerrada bajo la curva.



Precisión interpolada

La **precisión en *k*** y la **precisión media** se calculan para una única consulta; sin embargo, resulta mucho más interesante obtener una medida que involucre varias consultas. Para ello es necesario (1) establecer una serie de valores estándar para la exhaustividad (típicamente 0, 0.1, 0.2, …, 0.9 y 1.0), (2) transformar los valores de precisión para cada consulta a estos puntos estandarizados y (3) calcular el valor medio para todas las consultas en cada uno de los 11 puntos. Una vez obtenidos estos valores se pueden representar mediante una **curva de precisión-exhaustividad**.

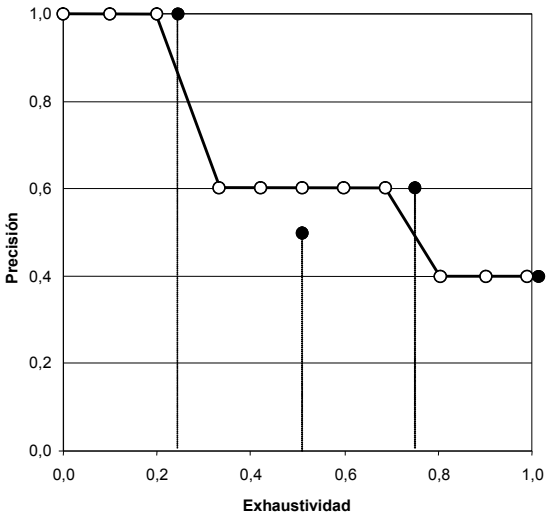
La precisión interpolada para un valor estándar de exhaustividad ρ no es más que el mayor valor de precisión para cualquier valor de exhaustividad experimental mayor o igual que ρ. Por ejemplo, dada la siguiente colección de documentos:

{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O}

Supongamos que para una consulta dada la lista de resultados es la siguiente:

{**E**, L, M, **I**, **O**, C, N, D, F, **A**, H, G, B, J, K}

Donde se muestran subrayados los documentos relevantes. Así pues, se tendría que para una exhaustividad de 0.25 la precisión es de 1; para 0.5 es 0.5; para 0.75 es 0.6 y para 1 es 0.4. Las precisiones interpoladas para los valores estándar de exhaustividad serán entonces: 1 para 0, 0.1 y 0.2; 0.6 para 0.3, 0.4, 0.5, 0.6 y 0.7 y 0.4 para 0.8, 0.9 y 1. En la figura se muestran los valores de precisión obtenidos experimentalmente como círculos negros y los valores interpolados como círculos blancos; también se muestra la curva de precisión-exhaustividad.



Precisión media (no interpolada)

La precisión media es una medida única del rendimiento de un sistema de **recuperación de información** que combina **precisión**, **exhaustividad** y la calidad de la ordenación de los resultados. Se trata de la media de los distintos valores de precisión para cada uno de los documento relevantes de la colección.

Por ejemplo, dada la siguiente colección de documentos:

{A, B, C, D, E, F, G, H, I, J, K, L, M, N, O}

Supongamos que para una consulta dada la lista de resultados es la siguiente:

{**E**, L, M, **I**, **O**, C, N, D, F, **A**, H, G, B, J, K}

Donde se muestran subrayados los documentos relevantes. Para calcular la precisión media es necesario calcular la precisión cada vez que se recupera un documento relevante hasta haber completado todos los documentos relevantes. Así, los valores de precisión serían: 1 al recuperar el documento E; 0,5 al recuperar el documento I; 0,6 al recuperar el documento O y 0,4 al recuperar el documento A. Por tanto la precisión media sería:

$$\frac{1+0{,}5+0{,}6+0{,}4}{4} = \frac{2{,}5}{4} = 0{,}625$$

Dicho de otro modo:

$$precisionMedia = \frac{\sum_{r=1}^{N} P(r) \cdot rel(r)}{R}$$

Donde N es el número de resultados; r es el resultado r-ésimo; $P(r)$ es la **precisión en *r***; $rel(r)$ es 1 si el resultado r-ésimo es relevante para la consulta y 0 en caso contrario y R es el número de documentos relevantes para la consulta.

Procesamiento de Lenguaje Natural (PLN)

El Procesamiento de Lenguaje Natural (PLN) es el conjunto de técnicas algorítmicas que tienen como objeto la manipulación y generación de muestras de lenguaje humano tanto en su manifestación escrita como oral. Ejemplos de técnicas de PLN son la generación de habla a partir de texto, el reconocimiento del habla, la traducción automática o la **recuperación de información**.

Programación genética

La programación genética es una técnica para la generación automática de programas de ordenador que alcancen el mejor rendimiento posible en una tarea definida por el usuario: generalmente aproximar una función desconocida pero para la cual se conoce su comportamiento deseado para un conjunto de datos de entrada. En su forma más sencilla los programas son simples expresiones representadas en forma de árbol.

Pureza

La pureza es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) mediante la comparación con una clasificación previa (que proporciona una serie de clases). No es más que la proporción entre el número de *ítems* pertenecientes a la clase dominante en un grupo y el tamaño de dicho grupo. Es decir, la pureza evalúa en qué medida un grupo de una clasificación automática contiene elementos de una única clase.

Recall

Véase **exhaustividad**.

Recomendación por contenidos

La recomendación por contenidos es una técnica que permite a un sistema proporcionar documentos similares a un documento de partida y que precisa, por tanto, de algún tipo de medida de **similitud** entre documentos.

Recuperación de información (*IR* o *information retrieval*)

El término recuperación de información hace referencia, en general, al estudio de sistemas automáticos que permitan a un usuario determinar la existencia o inexistencia de documentos (esto es, textos) relativos a una necesidad de información formulada habitualmente como una **consulta**.

Red neuronal

Técnica de **categorización** consistente en una estructura de capas interconectadas y formadas por elementos de procesamiento cuya funcionalidad está inspirada en las neuronas animales. Las redes neuronales requieren al menos dos capas: una de entrada con tantos elementos como variables definan a los objetos del problema y otra de salida con tantos elementos como categorías deba reconocer la red neuronal. Opcionalmente puede haber una o más capas ocultas. El aprendizaje de la red se produce mediante un entrenamiento durante el cual se ajustan los pesos de los distintos nodos de la red.

Reglas de decisión

Una técnica de categorización, no necesariamente automática, similar en cierta medida a los **árboles de decisión** y que consiste en la utilización de un conjunto de reglas para categorizar algún tipo de objetos en función de una serie de variables que los definen. Por su propia naturaleza las reglas de decisión son susceptibles de ser producidas de manera manual por los propios usuarios (p.ej. para dirigir correo electrónico con determinadas palabras en el asunto a una carpeta en particular).

Relevancia

La relevancia es una medida de la **similitud** entre los contenidos de un documento y la **consulta** de un usuario. Se trata de un valor subjetivo y cambiante, por lo que el término no suele hacer referencia al "juicio" que emitiría un usuario sino al valor que un sistema de **recuperación de información** asigna a cada documento en relación con una consulta. El objetivo de tales sistemas es producir valores de relevancia próximos a los que asignaría el propio usuario.

Resumen abstractivo/extractivo (*Abstract vs. Extract*)

Según Hovy (1999) un resumen extractivo (*extract*) consiste en una selección de parte del material presente en un documento original mientras que un resumen abstractivo (*abstract*) consiste en una condensación y reformulación del original. Un resumen extractivo se considera, a su vez, **informativo** mientras que uno abstractivo suele ser **indicativo**. Por lo que respecta al **resumen automático** resulta mucho más sencillo producir resúmenes por extracción que por abstracción.

Resumen automático (*automatic summarization*)

Las técnicas de resumen automático tienen como misión obtener a partir de un documento o conjunto de documentos un único texto mucho más corto que aún contenga los aspectos más relevantes de los originales. Durante los años 1950 y 1960 la investigación en este tipo de tecnologías fue intensa para descender considerablemente durante los años siguientes y no recuperarse hasta los años 1990. Desde entonces se trata de un campo muy activo y aunque aún se está lejos de disponer de sistemas capaces de emular a un ser humano (p.ej. produciendo

resúmenes indicativos) se ha avanzado enormemente y los sistemas estadísticos y puramente **extractivos** (o de "cortar-y-pegar") han demostrado su gran utilidad.

Resumen indicativo/informativo

Según Hovy (1999) *"un resumen informativo refleja el contenido del texto original, probablemente detallando sus argumentos, mientras que un resumen indicativo simplemente proporciona una indicación sobre el tema que trataba el documento original"*. Así, los resúmenes informativos suelen reemplazar a los documentos que resumen mientras que los indicativos permiten a los usuarios decidir sobre la pertinencia del documento en relación a una necesidad de información específica.

Ro (P)

En el contexto de la técnica ***blindLight*** se define P como el cociente de la **significatividad total** del vector **intersección** de una **consulta** y un documento entre la significatividad total de un vector documento. P revela en qué medida el documento "satisface" a la intersección entre éste y una consulta. Esta medida puede combinarse con **Pi (Π)** para construir distintas medidas de **similitud** adaptadas a las necesidades de las diferentes aplicaciones.

Rocchio, algoritmo de

Algoritmo para la expansión de **consultas** por realimentación (*relevance feedback*) que también se ha empleado como técnica de **categorización automática**. La idea subyacente a la técnica es sencilla: (1) dada una consulta, un sistema de recuperación de información proporciona al usuario un conjunto de documentos, (2) el usuario selecciona los que considera relevantes y (3) se "enriquece" la consulta original calculando la diferencia entre los documentos relevantes (POS_i, véase la ecuación) y los no relevantes (NEG_i). Así, una categoría c_i estaría representada por un vector de pesos w_{ki} calculados según la siguiente fórmula en la que w_{kj} es el peso del término t_k en el documento d_j.

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

Self-Organizing Maps (SOM)

Los Mapas Auto-Organizativos (Kohonen 1982) consisten en una **red neuronal** (generalmente bi o tridimensional) que se entrena con vectores de características en un proceso competitivo. Para cada vector hay una única neurona ganadora que ajustará sus pesos para aproximarse al vector de entrada. No obstante, el resto de neuronas también ajustan parcialmente sus pesos de forma inversamente proporcional a la distancia a que se encuentren de la vencedora. De este modo, se van vinculando los vectores a diferentes coordenadas del mapa y en caso de que estén etiquetados se asociarán sus etiquetas a las distintas zonas del mismo.

Significatividad

En el contexto de la técnica ***blindLight*** se denomina significatividad a los pesos asignados a cada uno de los *n*-gramas que forman un vector documental. Dichos pesos son calculados a partir tan sólo del texto original empleando alguno de los estadísticos propuestos por Ferreira da Silva y Pereira Lopes (1999) como la información mutua o la probabilidad condicional simétrica.

Significatividad total

En el contexto de la técnica ***blindLight*** la significatividad total es la suma de los valores de **significatividad** para todos los *n*-gramas que componen un documento.

Similitud

La similitud es una cantidad que refleja la fuerza de la asociación (o parecido) entre dos objetos en función de sus características. Suele variar en el rango [-1, 1] aunque puede normalizarse entre 0 y 1. La similitud normalizada puede transformarse de manera trivial en una **disimilitud** normalizada, si s_{ij} es la similitud normalizada entre los elementos *i* y *j* entonces la disimilitud entre ambos será:

$$\delta_{ij} = 1 - s_{ij}$$

En el **modelo vectorial** se emplea habitualmente la **función del coseno** para calcular la similitud entre documentos.

Similitud promedio (*overall similarity*)

La similitud promedio es una medida para la evaluación de soluciones de **agrupamiento** (que producen grupos de elementos) que no precisa de la comparación con una clasificación previa. Se trata tan sólo de la **similitud** media entre cada par de documentos de un grupo.

Stemming (reducción a la raíz)

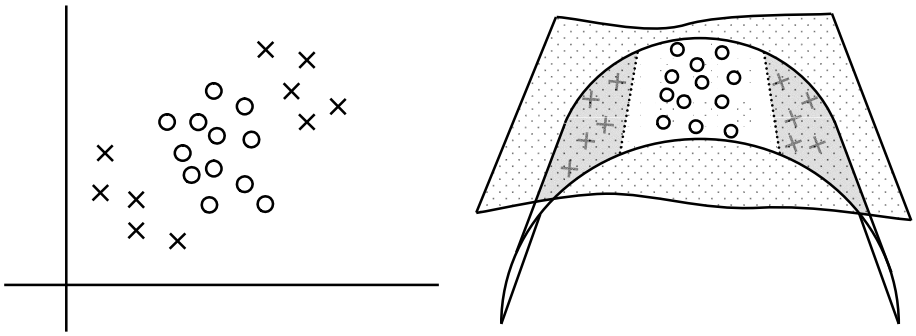
Un algoritmo de *stemming* o *stemmer* determina la raíz morfológica de una palabra colapsando múltiples formas de la raíz en un único término (`research`, `researcher`, `researching` y `researchers` colapsan en `research` empleando un *stemmer* para inglés). Un *stemmer* para castellano, por ejemplo, transformaría `andanzas` en `and`, `habitaciones` en `habit` o `jugueis` en `jug`.

Stop words

Se denominan *stop words* o palabras vacías aquellas palabras que, a pesar de un uso frecuente, aportan por sí solas poco significado a un texto. En la sentencia anterior se muestran subrayadas algunas palabras vacías del castellano.

Support Vector Machines (*SVM*)

Método de **categorización automática** propuesto por Boser, Guyon y Vapnik (1992) y que consiste en la transformación de los vectores de entrada (véase **modelo vectorial**), que definen una dimensión en la cual no son linealmente separables, a una dimensión superior que permita su separación mediante una única (hiper)superficie. Se trata de una de las técnicas de categorización más eficientes aunque, tal vez debido a su complejidad, otras que ofrecen peores resultados (como los **categorizadores bayesianos**) continuan siendo muy populares.



*tf*idf*

Método para ponderar los términos de un documento en un sistema de **recuperación de información**. Según este método la importancia de un término es directamente proporcional a su frecuencia de aparición en un documento (*tf*) e inversamente proporcional al número de documentos en que aparece (***idf***).

Valoración explícita/implícita

En los sistemas de **filtrado colaborativo** se recomiendan nuevos elementos a un usuario en función de sus preferencias pasadas y de la utilidad que dichos elementos para otros usuarios con preferencias similares. Para determinar dicha utilidad es necesaria una valoración por parte del usuario; dicha valoración puede ser explícita (p.ej. puntuando el elemento) o implícita, es decir, sin requerir la intervención del usuario y basándose tan sólo en su interacción con el elemento (p.ej. tiempo de lectura, impresión del documento, incorporación a la lista de enlaces favoritos, etc.)

Web Semántica

Según Tim Berners-Lee, James Hendler y Ora Lassila (2001) *"la Web Semántica es una extensión de la Web actual en la cual se asigna a la información un significado bien definido, posibilitando una mejor cooperación entre máquinas y usuarios"*. Una pieza clave en el desarrollo de la Web Semántica son las **ontologías**.