Procesamiento del Lenguaje Natural, num. 32 (2004), pp.67-74 recibido 6-12-2003; aceptado 20-2-2004

Fundamentos de Latent Semantic Indexing (LSI) y su aplicación a la categorización de textos periodísticos en euskara*

A. Zelaia Jauregi
Dpto. Ciencias de la Computación e Inteligencia Artificial
Facultad de Informática, Universidad del País Vasco UPV-EHU
ccpzejaa@si.ehu.es

Resumen: Muchos métodos de búsqueda de textos en Internet dependen de un emparejamiento exacto entre palabras que busca el usuario y las que existen en el documento. La descomposición en valores singulares utilizada por LSI permite recuperar información basada en conceptos o significados que están latentes en el documento. En este artículo se analizan los fundamentos matemáticos de dicha técnica, y se muestran unos resultados obtenidos para un experimento de categorización de textos. Además, se citan algunas aplicaciones de LSI para el procesamiento del lenguaje natural.
Palabras clave: Descomposición en valores singulares, latent semantic indexing.

Abstract: Currently, most approaches to retrieving textual materials from scientific databases depend on a lexical match between words in user's requests and those in documents in a database. Using the singular value decomposition, LSI takes advantage of the implicit higher-order structure in the association of terms with documents. Here we describe the mathematical foundations of this technique, and show the results that we have obtained by applying it to a text categorization experiment. Morover, we note the applications of LSI in NLP.
Keywords: Singular value decomposition, latent semantic indexing.

1 Introducción.

Latent Semantic Indexing (LSI), que surgió en 1988, es una teoría y un método para la extracción y representación del conocimiento (Dumais et al., 1988) (Deerwester et al., 1990). Se basa en la utilización contextual de las palabras en un corpus de gran tamaño para extraer y representar el significado de las palabras y conjuntos de palabras, haciendo uso de cálculos estadísticos. LSI surgió como una herramienta para la indexación y recuperación automática de la información, y con el propósito de superar una deficiencia fundamental de algunas técnicas de recuperación de la información: el problema de la semántica.

Fue unos años más tarde cuando un grupo de psicólogos, mediante una serie de experimentos, llegaron a considerar que LSI refleja de forma adecuada el conocimiento y aprendizaje humano. Se quiso dar respuesta a uno de los misterios más persistentes de la cognición; aquel que cuestiona cómo conseguirán los seres humanos obtener tanto conocimiento estando expuestos a tan poca información. Dado que LSI, partiendo de textos escritos y utilizando la técnica de la descomposición en valores singulares de una matriz es capaz de captar las inter-relaciones existentes entre términos y documentos, se pensó realizar un experimento para comprobar hasta qué punto LSI es capaz de simular el aprendizaje de lenguaje humano. Así, LSI fue entrenado utilizando un corpus de aproximadamente 5 millones de palabras de textos enciclopédicos. A continuación fue sometido a la prueba TOEFL (Test Of English as a Foreign Language). Dicha prueba consiste en mostrar una palabra o frase corta y pedir que se seleccione una de entre las cuatro opciones dadas que más se le asemeje en significado. LSI respondió con acierto en el 64 % de los casos; tasa de acierto que coincide con la que obtienen

* Agradecimientos: Los experimentos descritos se enmarcan en el proyecto "Aplicación de técnicas basadas en el aprendizaje automático para la clasificación y recuperación de documentos escritos en euskara" que ha sido parcialmente subvencionado por el Gobierno Vasco dentro del programa Universidad-Empresa (Código UE02/B11), por la Universidad del País Vasco UPV00141.226-T-14816/2002 y por la Diputación de Gipuzkoa dentro de un programa de la Comunidad Europea.

ISSN 1135-5948

los estudiantes de países de habla no inglesa que pasan la prueba para acceder a institutos de Estados Unidos. Esto les llevó a proponer que LSI consigue aprender a medir la similitud semántica entre palabras basándose solamente en texto escrito de forma similar a la que hacen los humanos (Landauer y Dumais, 1997).

Analizar la similaridad semántica entre palabras puede ser un buen método para medir el conocimiento de una lengua. Sin embargo, no es suficiente para evaluar la correspondencia entre el conocimiento que obtiene LSI a partir de textos y el que obtienen los seres humanos, ya que a menudo los seres humanos expresan su conocimiento utilizando frases, párrafos o incluso fragmentos más grandes de texto. LSI ha sido utilizado con éxito para evaluar la calidad de textos escritos por alumnos en un entorno educacional, donde se ha comprobado que existe una fuerte correlación entre la evaluación realizada por LSI y la realizada por los expertos humanos (Foltz, Laham, y Landauer, 1999).

Desde entonces LSI ha sido utilizado en diversas aplicaciones que serán comentadas en el apartado 4. Cabe destacar el capítulo recopilatorio dedicado a LSI en el que se puede encontrar una amplia bibliografía (Dumais, 2004). Existen, además, en la web varias páginas con recursos disponibles. En la página LSI de Telcordia, http://lsi.research.telcordia.com, se recopilan artículos, demostraciones y software. En la página LSA de la Universidad de Colorado, http://lsa.colorado.edu, hay disponibles varias demostraciones, incluída una herramienta para la evaluación de redacciones y herramientas para el análisis de términos y frases. La página LSI de la Universidad de Tennessee http://www.cs.utk.edu/~lsi, contiene artículos, corpora y software para el análisis de textos y algoritmos para realizar de forma eficiente la descomposición en valores singulares.

En este artículo pasamos a estudiar los fundamentos de LSI tanto de forma teórica como a través de un ejemplo. A continuación se citan las aplicaciones de LSI en el procesamiento del lenguaje natural y finalmente se recogen los resultados de un experimento de categorización de textos utilizando LSI.

2 Fundamentos de LSI.

En este apartado se analiza la base matemática en la que se sustenta LSI de forma superficial.

2.1 Construcción de la matriz.

LSI comienza por construir una matriz $\mathbf{M}$ de *términos* por *documentos* para un corpus dado (Berry y Browne, 1999), donde la posición m_{ij} de la matriz representa la frecuencia con la que aparece el término i en el documento j .

Antes de realizar la descomposición en valores singulares de la matriz, se da la posibilidad de realizar un preprocesado utilizando pesos. Así es como se transforma la matriz de frecuencias $\mathbf{M}$ en una matriz de pesos:

$$m'_{ij} = L(i,j) \cdot G(i).$$

Se pueden utilizar los siguientes pesos:

- Peso local $L(i,j)$. Mide la importancia del término i en el documento j. Se puede mantener la matriz de frecuencias, convertirla a binaria o reducir las diferencias entre frecuencias utilizando la función de logaritmo.
- Peso global $G(i)$. Mide la importancia del término i a nivel de corpus. Se tienen las siguientes opciones: dar la misma importancia a todos los términos, normalizar los vectores que representan a los términos, calcular el **idf** o calcular la entropía. En general, los pesos globales asocian valores más bajos a términos que aparecen muy frecuentemente o en muchos documentos.

Se suele recomendar utilizar como peso local el logaritmo y como peso global la entropía porque es la combinación que mejores resultados da (Dumais, 1991).

2.2 Descomposición en valores singulares de la matriz

Un resultado muy conocido en álgebra asegura que toda matriz rectangular $\mathbf{M} \in \mathbb{R}^{m \times n}$ puede ser escrita como producto de otras tres matrices

$$\mathbf{M} = \mathbf{U}\Sigma\mathbf{V}^T = \sum_{i=1}^{k} \sigma_i u_i v_i^T$$

donde $\Sigma \in \mathbb{R}^{m \times n}$ es una matriz diagonal de valores singulares $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_k \geq 0$

siendo $k = \min\{m, n\}$, y las matrices $\mathbf{U}$ y $\mathbf{V}$ son matrices formadas por vectores singulares (Golub y Loan, 1996). Los valores singulares son únicos, y el número de valores singulares distintos de cero es el rango r de la matriz $\mathbf{M}$, siendo $r \leq k$.

Dada una matriz, existen técnicas para calcular los valores singulares y sus correspondientes vectores singulares (Berry et al., 1993), y así es como LSI calcula las matrices $\mathbf{U}$, Σ y $\mathbf{V}$.

2.3 Reducción de la dimensión del espacio.

Una vez que se dispone de una descomposición en valores singulares de la matriz $\mathbf{M}$, es posible aproximarla por otra $\mathbf{M}_p$ de rango $p \leq r$, calculada para los primeros p valores singulares.

$$\mathbf{M}_p = \mathbf{U}_p \Sigma_p \mathbf{V}_p^T = \sum_{i=1}^{p} \sigma_i u_i v_i^T$$

donde $\mathbf{U}_p$ y $\mathbf{V}_p$ son matrices formadas por las p primeras columnas de las matrices $\mathbf{U}$ y $\mathbf{V}$. Se asegura que esta matriz $\mathbf{M}_p$ es una de las matrices de rango p o menor que mejor se aproxima a la matriz original $\mathbf{M}$,

$$\min_{rang(A) \leq p} ||\mathbf{M} - \mathbf{A}|| = ||\mathbf{M} - \mathbf{M}_p|| = \sigma_{p+1}.$$

Como consecuencia de esta operación, se pasa del espacio vectorial generado por las columnas de la matriz $\mathbf{M}$ al espacio generado por las columnas de $\mathbf{M}_p$. Esta operación se dice *reducción de la dimensión* y a este espacio de dimensión p se le llama **espacio reducido**.

¿Es válida esta aproximación? Dicho de otro modo, ¿por qué la aproximación de $\mathbf{M}$ por $\mathbf{M}_p$ funciona desde un punto de vista semántico? Los diversos autores hablan aquí de que la variación en el uso del vocabulario y la ambigüedad de muchas palabras producen *ruido* significativo en $\mathbf{M}$. Al tomar $\mathbf{M}_p$ en lugar de $\mathbf{M}$ se captura lo suficiente de la estructura que asocia términos y documentos para retener su significado *oculto*, y por tanto, se consigue quitar *ruido*.

El usuario debe decidir el rango p para el que se aproximará la matriz, es decir, la dimensión del espacio reducido. Esta es una decisión fundamental, ya que del acierto con el que se elija p dependen fuertemente los resultados. Sin embargo, no se conoce ningún método para seleccionar la p adecuada, y por eso, su elección se suele basar en la experimentación (Dumais, 1991) (Letsche y Berry, 1997). Se recomienda seleccionar un valor entre 100 y 300.

2.4 Búsqueda semántica.

Ahora que se tiene generado el espacio reducido es cuando se pueden realizar los cálculos de similitud semántica con los vectores de dicho espacio. Desde el punto de vista de la extracción de la información, al introducir una petición en el buscador, se considera que se ha dado un **vector de búsqueda**

$$\mathbf{q}^T = (q_1, q_2, \ldots, q_m)$$

que LSI construye observando si cada término i de su base de datos aparece ($q_i = 1$) o no ($q_i = 0$) en la petición de búsqueda. Se aplica el preprocesado mediante pesos a este vector, y a continuación se traslada el vector $\mathbf{q}$ al espacio reducido, de la misma manera que se ha procedido con los documentos que forman el corpus.

$$\mathbf{q}_p = \mathbf{q}^T \mathbf{U}_p \Sigma_p^{-1}.$$

Las componentes del vector $\mathbf{q}_p$ son las coordenadas de la proyección del vector $\mathbf{q}$ en el espacio reducido.

A continuación se calcula la similaridad que existe entre el vector de búsqueda $\mathbf{q}_p$ y los vectores $\mathbf{d}_p$ que representan a los documentos del corpus. Como medidas de similitud se pueden tomar tanto el producto escalar entre dos vectores como el coseno del ángulo θ que forman. Se recomienda utilizar el coseno porque da mejores resultados.

$$cos\ \theta = \frac{\mathbf{q}_p^T \mathbf{d}_p}{||\mathbf{q}_p|| ||\mathbf{d}_p||}$$

Si el ángulo θ que forman los dos vectores es pequeño, el coseno será próximo a 1, y se interpretará que la búsqueda y el documento son semánticamente muy similares. Aquellos documentos cuyo coseno sea superior a un umbral serán devueltos al usuario.

3 Ejemplo.

Para facilitar la compresión de la teoría expuesta en el apartado anterior, se va a mostrar un pequeño ejemplo hallado en la bibliografía (ver (Landauer, Laham, y Foltz, 1998), (Landauer y Dumais, 1997) y (Deerwester et al., 1990)). El corpus está formado por los siguientes 9 documentos:

c1: Human machine interface for Lab ABC computer applications.

c2: A survey of user opinion of computer system response time.

c3: The EPS user interface management system.

c4: System and human system engineering testing of EPS.

c5: Relation of user-perceived response time to error measurement.

m1: The generation of random, binary, unordered trees.

m2: The intersection graph of paths in trees.

m3: Graph minors IV: Widths of trees and well-quasi-ordering.

m4: Graph minors: A survey.

Los documentos c1-c5 son títulos de publicaciones sobre "interacción entre persona-computador" y los documentos m1-m4 tratan sobre la "teoría de grafos". Las palabras subrayadas son los términos elegidos por LSI, 12 en total. Se construye la siguiente matriz $\mathbf{M} \in \mathbb{R}^{12\times 9}$.

	c1	c2	c3	c4	c5	m1	m2	m3	m4
	↓	↓	↓	↓	↓	↓	↓	↓	↓
human	1	0	0	1	0	0	0	0	0
interface	1	0	1	0	0	0	0	0	0
computer	1	1	0	0	0	0	0	0	0
user	0	1	1	0	1	0	0	0	0
system	0	1	1	2	0	0	0	0	0
response	0	1	0	0	1	0	0	0	0
time	0	1	0	0	1	0	0	0	0
EPS	0	0	1	1	0	0	0	0	0
survey	0	1	0	0	0	0	0	0	1
trees	0	0	0	0	0	1	1	1	0
graph	0	0	0	0	0	0	1	1	1
minors	0	0	0	0	0	0	0	1	1

Por simplicidad se eligen los pesos que mantienen la matriz de frecuencias. A continuación, se calcula la descomposición en valores singulares y se elige $p = 2$, con lo que se obtiene la siguiente matriz de aproximación $\mathbf{M}_2$.

$$\begin{pmatrix}
0.16 & 0.40 & 0.38 & 0.47 & 0.18 & -0.05 & -0.12 & -0.16 & -0.09 \\
0.14 & 0.37 & 0.33 & 0.40 & 0.16 & -0.03 & -0.07 & -0.10 & -0.04 \\
0.15 & 0.51 & 0.36 & 0.41 & 0.24 & 0.02 & 0.06 & 0.09 & 0.12 \\
0.26 & 0.84 & 0.61 & 0.70 & 0.39 & 0.03 & 0.08 & 0.12 & 0.19 \\
0.45 & 1.23 & 1.05 & 1.27 & 0.56 & -0.07 & -0.15 & -0.21 & -0.05 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.16 & 0.58 & 0.38 & 0.42 & 0.28 & 0.06 & 0.13 & 0.19 & 0.22 \\
0.22 & 0.55 & 0.51 & 0.63 & 0.24 & -0.07 & -0.14 & -0.20 & -0.11 \\
0.10 & 0.53 & 0.23 & 0.21 & 0.27 & 0.14 & 0.31 & 0.44 & 0.42 \\
-0.06 & 0.23 & -0.14 & -0.27 & 0.14 & 0.24 & 0.55 & 0.77 & \boxed{0.66} \\
-0.06 & 0.34 & -0.15 & -0.30 & 0.20 & 0.31 & 0.69 & 0.98 & 0.85 \\
-0.04 & 0.25 & -0.10 & -0.21 & 0.15 & 0.22 & 0.50 & 0.71 & 0.62
\end{pmatrix}$$

Utilizando correlaciones de Pearson, se puede interpretar la transformación de la matriz $\mathbf{M}$ en $\mathbf{M}_2$ de la siguiente manera.

(a) Transformación de los valores m_{ij} de la matriz. Observemos el término trees en el documento m4 tanto en la matriz $\mathbf{M}$ como en la $\mathbf{M}_2$ (ver valores recuadrados). Dicho término no aparece en el documento m4, razón por la que la frecuencia de aparición es 0. Sin embargo, en la matriz $\mathbf{M}_2$ pasa a ser 0.66, lo que refleja de forma más adecuada la importancia que el término trees tiene en la teoría de grafos. Esta transformación ha ocurrido gracias a otros términos que aparecen en el contexto del término trees como son graph y minors.

(b) Tranformaciones a nivel de términos. Observemos los vectores que representan a los términos human y user (filas 1 y 4 de las matrices). Se puede apreciar que ningún documento del corpus contiene ambos términos, y por eso, la correlación entre dichos vectores en $\mathbf{M}$ nos da negativa (r=-0.38). Sin embargo, es evidente que los términos "humano" y "usuario" son semánticamente cercanos. Al calcular la correlación sobre los vectores de la matriz $\mathbf{M}_2$ se obtiene un valor positivo cercano a 1 (r=0.94).

(c) Transformaciones a nivel de documentos. Se ha calculado la correlación entre todo par de vectores columna. En el Cuadro 1 se muestran las medias de dichas correlaciones.

	M		**$\mathbf{M_2}$**	
	c1-c5	m1-m4	c1-c5	m1-m4
c1-c5	0.02	-0.30	0.92	-0.72
m1-m4	-0.30	0.44	-0.72	1.00

Tabla 1: Medias de correlaciones entre pares de documentos.

Se observa que en $\mathbf{M}$ las únicas correlaciones positivas son las calculadas al comparar documentos del mismo tema. En $\mathbf{M}_2$, además de ser positivas se acercan a 1, con lo se muestra más claramente la similitud semántica entre ellos.

Por todo ello se concluye que la reducción de la dimensión ha conseguido realizar muchas inferencias semánticas apropiadas.

3.1 Representación gráfica de los documentos en el espacio reducido.

Dado que se ha reducido la dimensión del espacio a $p = 2$, es posible representar gráficamente los documentos en el plano. Las coordenadas en el plano de la proyección de cada uno de los 9 vectores $\mathbf{d}$ las calculamos de la siguiente manera.

$$\mathbf{d}_2 = \mathbf{d}^T \mathbf{U}_2 \Sigma_2^{-1}.$$

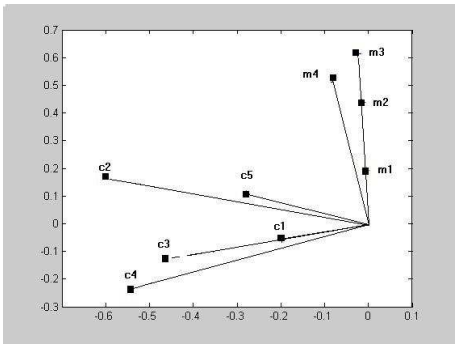

Figura 1: Representación de los documentos en el plano.

En la Figura 1 se observa que el ángulo entre vectores del mismo tema es pequeño, con lo que el coseno se acerca a 1, y se consideran semánticamente similares.

De la misma manera se pueden calcular y representar los 12 términos, así como la búsqueda introducida por el usuario.

4 Aplicaciones en el área del PLN.

Cada vez son más las aplicaciones de LSI, tanto en el área pedagógico/cognitiva como en el procesamiento del lenguaje natural. Entre estas últimas podemos destacar las siguientes:

Categorización de documentos: En la sección 5 se expone detalladamente esta utilización.

Recuperación multilingüe: Se obtienen buenos resultados utilizando "documentos duales" donde cada documento es el conjunto de palabras correspondiente a la redacción del mismo texto en distintos idiomas. Con ello se consigue no hacer ningún tipo de traducción ya que por medio de la descomposición en valores singulares se relacionan términos en distintos idiomas. Los resultados son comparables a los obtenidos utilizando sistemas de traducción automática, con un coste mucho menor. (Dumais et al., 1997)

Evaluación automática de resúmenes y análisis del contenido y coherencia de textos: Los sistemas de evaluación de resúmenes y redacciones basadas en LSI se centran en medir su contenido conceptual. Para ello, se entrena LSI utilizando un conjunto de textos representativos del dominio. Se incluyen en este espacio semántico redacciones evaluadas y se compara con ellas la que se pretende evaluar. (Landauer et al., 1997), (Kintsch et al., 2000), (Wiemer-Hastings et al., 1998) (Foltz, Kintsch, y Landauer, 1998).

Modelado del lenguaje en reconocimiento del habla: Se utiliza una representación vectorial basada en LSI como complemento a la convencional basada en trigramas para beneficiarse del conocimiento semántico que incluye el modelo LSI (Bellegarda, 2000) (Deng y Khudanpur, 2003).

Desambigüación del sentido de las palabras: Dentro de un algoritmo no supervisado se inducen las similitudes entre las palabras basándose en co-ocurrencias de términos con palabras de su contexto. La dimensión de este espacio es reducida usando la descomposición en valores singulares. Finalmente usando técnicas de clustering se identifican los sentidos de las palabras. (Schutze, 1998) (Widdows y Peters, 2003).

5 Aplicación a la categorización de documentos.

Uno de los ámbitos en los que LSI ha sido utilizado con éxito es en la categorización de documentos escritos (Dolin et al., 1999) (Sebastiani, 2002) (Rosso et al., 2004) (Pierre, 2001). La categorización de documentos consiste en etiquetar los textos escritos en lenguaje natural con una categoría elegida de entre un conjunto de categorías temáticas previamente establecido. La categorización se realiza en dos fases: la fase de *entrenamiento* en la que se obtiene una generalización inductiva del conjunto de documentos que se utilizan para el aprendizaje del sistema, y la de *test* que se encargará de evaluar la efectividad del mismo. Para ello, es necesario disponer de un conjunto de textos clasificados manualmente.

El experimento realizado pretende categorizar documentos escritos en euskara, lengua de rica flexión y aglutinante. Debido a esto, en los textos escritos en esta lengua encon-

tramos muchas veces formas similares con un lema o raíz común. El sufijo de este lema es el que va a hacer que dos palabras sean diferentes, bien en número, en determinación y en caso para los sustantivos, o bien en modo, tiempo, aspecto, persona y número en el caso de las formas verbales. Esta cuestión resulta importante para el problema de clasificación que nos ocupa; lo que realmente contiene información semántica no es tanto la palabra sino el lema que le corresponde. En un sistema de este tipo, cuanto mayor sea la frecuencia de aparición de esa palabra en un texto, mayor será el peso que se le asignará, y si su frecuencia no llega a un mínimo suele ser descartada. Por tanto, en el caso de basarnos en palabras, aunque un mismo lema se repita en numerosas ocasiones, al estar presente en diferentes formas, puede llegar a no considerarse. En consecuencia, la lematización puede ser un proceso importante en cuanto que reduce la dimensión de la información a tratar e incluso puede producir una mejora en la eficiencia del sistema (Alegria et al., 2004). Para obtener los lemas correspondientes a cada palabra, se ha utilizado la herramienta diseñada por el grupo IXA[1]. En (Ezeiza et al., 1998) se presentan las características de este lematizador.

El corpus que se ha utilizado en este trabajo proviene del diario *Euskaldunon Egunkaria*. Se parte de un conjunto de 6.064 documentos[2] etiquetados (Arregi y Fernández, 2002). Este corpus ha sido repartido en dos colecciones: 4.548 documentos para aprendizaje y 1.516 para test. Además, se ha procedido a lematizar el corpus, con lo que se dispone de dos corpus de aprendizaje y dos de test, para poder analizar el impacto que la lematización tiene en la categorización de textos en euskara.

Todos estos documentos llevan asociada originalmente una única de las siguientes 7 categorías: economía, europa, sociedad, deportes, cultura, mundo, política. Sin embargo, y con el objeto de adecuar el experimento a un proyecto más global, se han distribuído estos documentos a las 17 categorías estándares[3]. Este trasvase se ha realizado mediante un proceso de *bootstrap*, con una posterior revisión manual. Hay que resaltar, que la distribución de documentos por categorías no es uniforme, tal como se puede observar en el Cuadro 2.

Categoría	Docum. aprend.	Docum. test
Cultura	600	202
Justicia	129	42
Desastres	75	26
Economía	234	78
Educación	82	27
Medioambiente	69	22
Salud	35	12
Intereses humanos	36	11
Trabajo	132	43
Ocio	40	13
Política	1.184	393
Religión	25	8
Tecnología	35	12
Sociedad	464	156
Deportes	1.283	429
Guerras	100	33
Metereología	25	9
Total	4.548	1.516

Tabla 2: Número de documentos por categorías.

Una vez que se han tenido los corpus preparados, se ha procedido a realizar el experimento. Las palabras que aparecen en al menos dos documentos han sido seleccionadas como términos.

Para el corpus no lematizado se ha generado una matriz de 34.288 términos $\times$ 4.548 documentos. Se ha realizado la descomposición en valores singulares y se ha calculado el coseno de cada documento del test con cada uno de los de aprendizaje utilizando LSI. A continuación, se ha realizado su categorización en base a los más cercanos (de mayor coseno) según el algoritmo k-NN. Tal como se puede observar en el Cuadro 3, se han hecho diferentes pruebas ($p = 100, 200, 300, 400, 500$) para diferentes dimensiones (ver apartado 2.3). Para el corpus no lematizado (NL) los mejores resultados se han obtenido para $p = 300$ valores singulares utilizando $k = 10$ vecinos más cercanos, con lo que se ha conseguido una tasa de acierto del 84.9%.

Se ha repetido el experimento utilizando el corpus lematizado (L). En este caso la matriz generada ha sido de menor tamaño, ya que tan solo 14.648 términos (lemas) han sido seleccionados. Los mejores resultados se han

[1]Grupo IXA: http://ixa.si.ehu.es

[2]Proyecto Hermes, (TIC-2000-0335-C03-03)

[3]Primer nivel en IPTC, http://www.iptc.org

obtenido para $p = 400$ observando $k = 3$ vecinos más cercanos, consiguiéndose una tasa de acierto del 87.3%, sensiblemente mejor que con el corpus no lematizado.

		100	200	300	400
Micro-average	NL	83.00	84.30	84.90	84.80
	L	85.90	86.60	86.80	87.30

Tabla 3: Resultados de LSI combinado con k-NN.

Cabe resaltar que estos resultados son superiores a los obtenidos para el mismo corpus utilizando otras técnicas de aprendizaje como Naive Bayes (NB), Winnow y Support Vector Machines (SVM). Los datos que aparecen en el Cuadro 4 corresponden al caso en el que las palabras que aparecen en menos de 4 documentos han sido filtradas. Se tienen 17.776 características para el corpus no lematizado (NL) y 8.542 para el corpus lematizado (L).

		NB	Winnow	SVM
Micro-average	NL	77.77	79.49	83.71
	L	80.28	77.77	83.11

Tabla 4: Resultados para otros métodos.

Sin embargo, hay que decir que el proceso de clasificación utilizando LSI es lento, debido a que hay que calcular los cosenos con todos los documentos del corpus de aprendizaje. Dado que las ventajas de LSI estan en la descomposición en valores singulares y la reducción de la dimensión, a corto plazo vamos a utilizar otras técnicas de clasificación que no conlleven la comparación con todos los documentos. Con ello conseguiríamos una categorización más rápida, y quizás, de mayor precisión.

Bibliografía

Alegria, I., O. Arregi, I. Fernández, I. Fernández, y A. Zelaia. 2004. Lemmatisation for text categorization in basque. submited for publication.

Arregi, O. y I. Fernández. 2002. Clasificación de documentos escritos en euskara. impácto de la lematización. *I. Jornadas de Tratamiento y Recuperación de Información, JOTRI, Valencia*, páginas 29–35.

Bellegarda, J.R. 2000. Exploiting latent semantic information in statistical language modeling. En *Proceedings of the IEEE*, volumen 88, páginas 1279–1296, August.

Berry, M., T. Do, G. O'Brien, V. Krishna, y S. Varadhan. 1993. Svdpackc: Version 1.0 user's guide. Tech.Rep. CS-93-194, University of Tennessee, Knoxville, October.

Berry, M.W. y M. Browne. 1999. *Understanding Search Engines: Mathematical Modeling and Text Retrieval*. SIAM Society for Industrial and Applied Mathematics, ISBN: 0-89871-437-0, Philadelphia.

Deerwester, S., S.T. Dumais, G.W. Furnas, T.K. Landauer, y R. Harshman. 1990. Indexing by latent semantic analysis. *Journal of the American Society for Information Science*, 41:391–407.

Deng, Y. y S. Khudanpur. 2003. Latent semantic information in maximum entropy language models for conversational speech recognition. En *HLT-NAACL Human Language Technology Conference.*, páginas 56–63. Noth American Association for Computational Linguistics, Proceedings of the Main Conference, May-June.

Dolin, R., J. Pierre, M. Butler, y R. Avedon. 1999. Practical evaluation of ir within automated classification systems. En *Proceedings of the International Conference on Information and Knowledge Management CIKM*, páginas 322–329, Kansas city, Missouri, USA, November. ACM, ISBN:1-58113-146-1.

Dumais, S. 2004. Latent semantic analysis. En *ARIST (Annual Review of Information Science Technology)*, volumen 38, páginas 189–230.

Dumais, S.T. 1991. Improving the retrieval of information from external sources. *Behavior Research Methods, Instruments and Computers*, 23(2):229–236.

Dumais, S.T., G.W. Furnas, T.K. Landauer, S. Deerwester, y R. Harshman. 1988. Using latent semantic analysis to improve access to textual information. En *Proceedings of CHI'88 Conference on Human Factors in Computing*, páginas 281–285, New York. ACM.

Dumais, S.T., T.A. Letsche, M.L. Littman, y T.K. Landauer. 1997. Automatic

cross-language retrieval using latent semantic indexing. *AAAI Spring Symposium on Cross-Language Text and Speech Retrieval*, March.

Ezeiza, N., I. Aduriz, I. Alegria, J.M. Arriola, y R. Urizar. 1998. Combining stochastic and rule-based methods for disambiguation in agglutinative languages. En *COLING-ACL'98*, Montreal (Canada), August 10–14.

Foltz, P.W., W. Kintsch, y T.K. Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse Processes*, 25(2-3):285–307.

Foltz, P.W., D. Laham, y T.K. Landauer. 1999. The intelligent essay assessor: Applications to educational technology. *Interactive Multimedia Electronic Journal of Computer-Enhanced Learning*, 1(2).

Golub, G.H. y C.F. Van Loan. 1996. *Matrix Computations. 3rd.ed.* The Johns Hopkins University Press, ISBN:0-8018-5414-8.

Kintsch, E., D. Steinhart, G. Stahl, y LSA Research Group. 2000. Developing summarization skills through the use of lsa-based feedback. *Interactive Learning Environments*, 8(2):87–109.

Landauer, T.K. y S.T. Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of the acquisition, induction and representation of knowledge. *Psychological Review*, 104:211–240.

Landauer, T.K., D. Laham, y P.W. Foltz. 1998. Introduction to latent semantic analysis. *Discourse Processes*, 25:259–284.

Landauer, T.K., D. Laham, B. Rehder, y M.E. Schreiner. 1997. How well can passage meaning be derived without using word order? a comparison of latent semantic analysis and humans. En *Proceedings of the 19th Annual Meeting of the Cognitive Science Society*, páginas 412–417.

Letsche, T. y M. Berry. 1997. Large-scale information retrieval with latent semantic indexing. *Information Sciences*, 100:105–137.

Pierre, J.M. 2001. On the automated classification of web sites. *Linköping Electronic Articles in Computer and Information Science*, 6.

Rosso, P., E. Ferretti, D. Jiménez, y V. Vidal. 2004. Text categorization and information retrieval using wordnet senses. En P. Sojka K. Pala P. Smrz C. Fellbaum, y P. Vossen, editores, *GWC Proceedings*, páginas 299–304.

Schutze, H. 1998. Automatic word sense discrimination. *Computational Linguistics*, 24(1):97–123.

Sebastiani, F. 2002. Machine learning in automated text categorization. *ACM Computing Surveys*, 34(1):1–47, March.

Widdows, D. y S. Peters. 2003. Word vectors and quantum logic experiments with negation and disjunction. En R.T. Oehrle y J. Rogers, editores, *Proceedings of Mathematics of Language*, volumen 8, páginas 141–154.

Wiemer-Hastings, P., A.C. Graesser, D. Harter, y the TRG. 1998. The foundations and architecture of autotutor. En *Proceedings of the 4th International Conference on Intelligent Tutoring Systems*, páginas 334–343, Berlin, Germany. Springer-Verlag.