LA INDEXACIÓN SEMÁNTICA LATENTE EN LA RECUPERACIÓN DE INFORMACIÓN

INTRODUCCIÓN

En la actualidad encontramos múltiples modelos y propuestas para recuperar información en Internet, como motores de búsqueda, sociedades de información, catálogos virtuales, software especializado entre otras, el objetivo principal de estas, consiste en vincular la información que se encuentra latente en la red y bases de datos, con las necesidades de información del usuario de manera que logre establecer la mayor precisión y relevancia en la búsqueda realizada. Uno de estos modelos se denomina La Indización Semántica Latente (ISL), este artículo pretende describir detalladamente como aporta la ISL a la recuperación de información.

RESUMÉN

La Indización Semántica Latente (ISL), es un modelo alternativo que maneja la búsqueda de información mediante la indexación de términos, ubicándolos en un contexto semántico común, esto mediante cálculos matemáticos especializados, que dan como resultado la simulación del análisis realizado, normalmente por un ser humano (agrupación de varios criterios y términos semánticamente relacionados), con la capacidad de memoria y almacenamiento de una maquina.

PALABRAS CLAVES: Valor de Descomposición Singular (VLS)**,** Similitud Latente, Análisis Semántico Latente, Espacio semántico.

ABSTRACT:

Latent semantic indexing (LSI) is a model which deals with the information retrieval by indexing terms in common semantic contexts. This terms organization uses specialized Mathematic calculations which result in the simulated analysis generally done by a human being. (Human being organize different related critter and semantic terms) with the memory capacity and machine storage.

KEYS WORDS: Singular Decomposition Value (VSL), Latent Similarity, Latent Semantic Analysis, Semantic Space.

QUÉ ES ISL?

La indexación semántica latente es una teoría matemática y modelo de aplicación en los sistemas de recuperación de información, que permite determinar el uso y las relaciones de un término con un contexto, vinculando procesos matemáticos, valores de descomposición y extracción de contenidos.

Las relaciones de los términos se obtienen a partir de un análisis jerárquico del origen del mismo, estableciendo su significado, palabras subordinadas y relacionadas, proceso dado mediante la utilización de una herramienta

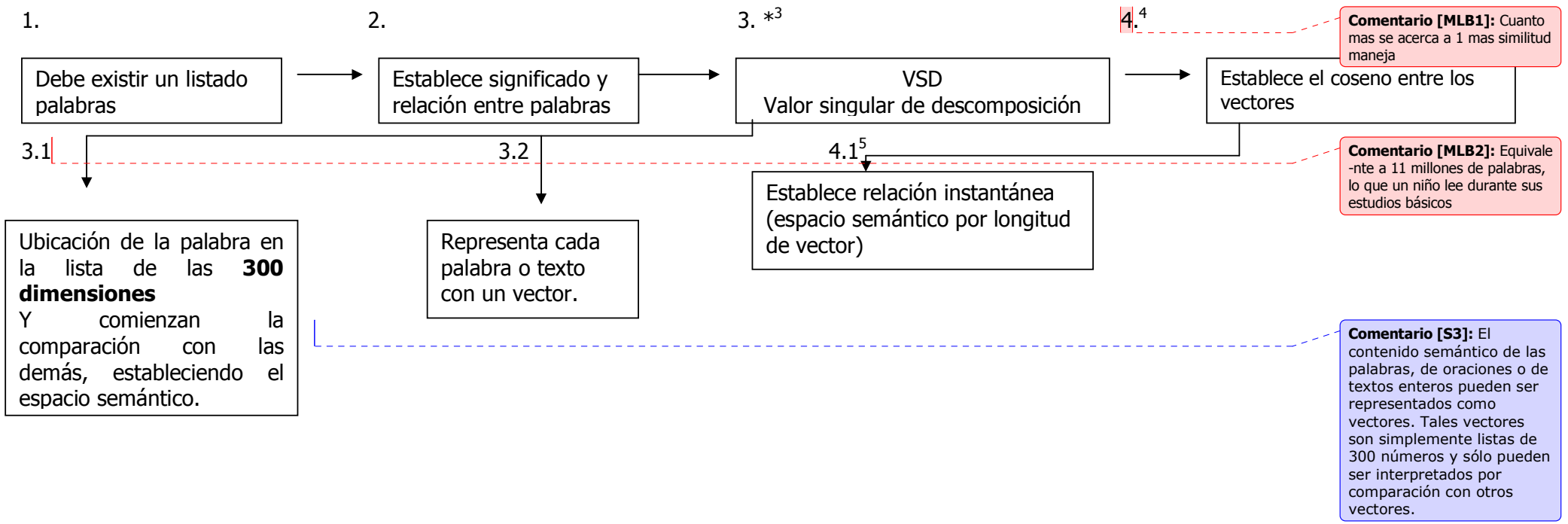
terminológica denominada tesauro[1] que permite la conversión del lenguaje natural (vocabulario manejado en la cotidianidad), y el lenguaje normalizado (Vocabulario estandarizado usado por los Bibliotecólogos), "La estructura de un tesauro se basa en vocabulario controlado y dinámico de términos que tienen entre ellos relaciones semánticas y genéricas y que se aplica a un dominio particular del conocimiento"[2].

Una vez se ha establecido un listado de palabras, se vinculan con un término dado (pregunta del usuario), se establece una matriz que se denomina de 300 dimensiones que es el listado de las mismas y, se les asigna un valor singular, con el fin de extraer aquellas más cercanas, e indexarlas con los documentos que las contienen mediante la obtención de un espacio semántico. En el gráfico que se encuentra a continuación se describe con mayor detalle y paso a paso el proceso de Indización semántica latente realizado por la maquina.

[1] Definición según P. Levery: **Tesauro:** Es el puente entre el lenguaje del informado (documentalista) y el lenguaje del no informado (usuario).

[2] Que es un tesauro : Disponible en http://web.usal.es/~alar/Bibweb/Temario/Tesauro.PDF Consultado el 05 de marzo de 2007 consultado en abril de de 2007

El proceso que realiza la maquina para la obtención de relaciones semánticas es de la siguiente manera.



[3] VSD: Técnica que utiliza un valor singular de descomposición que segmenta una gran matriz de datos de asociación de término- documento y permite construir un "espacio semántico" en el que se asocian entre sí términos y documentos. Venegas, René. Revista Signos. Documento LSI

[4] Coseno varia entre -1 y +1; 0 no existe similitud.

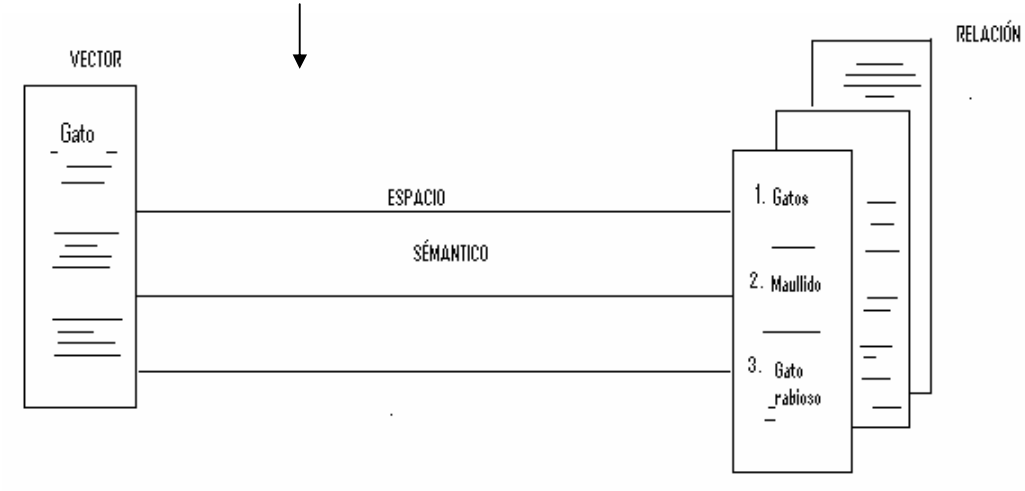
[5] La longitud del vector permite establecer que tanta experiencia maneja del término frase o párrafo a partir de la frecuencia de los mismos.

5.

A partir de la longitud de valor, realiza comparación de palabras arbitrarías con significado en la frase. Aprende la diferenciación en contextos.

6.

Lista las palabras que se encuentran cercanas a un vector



7.

A partir de los vectores encontrados, realiza una comparación frente a otros textos que manejan la misma palabra estableciendo similitud.

Para entender mejor este proceso, simularemos una búsqueda en Google: Por ejemplo, cuando buscamos la palabra "gato":

1. Primero establece el significado de "gato" (utilizando un diccionario de términos)
2. Analiza por medio del tesauro cuales son las palabras relacionadas, subordinadas y a que familia terminológica pertenece (obteniendo palabras como gatos, maullido, ratón, gato rabioso).
3. Con esta información se dirige a sus bases de datos e indexa los documentos que manejan el término y las palabras relacionadas.

La ISL depende de un poderoso análisis matemático que es capaz de inferir correctamente relaciones muy estrechas a partir de un análisis estadístico de las palabras en uso, estableciendo "verdadera representación semántica, un espacio que captura las relaciones semánticas esenciales" (Kintsch ,2002:5).En la tabla 1 (a continuación), podemos observar las relaciones que se encuentran frente a la palabra "gato", recordemos que entre mas cerca se encuentre de 1 mas acercamiento en cuanto a similitud, y por tanto mas cerca de la respuesta a la pregunta formulada por el usuario (vectores que se encuentran cerca).
En la columna 1 encontramos el numero del ítem o palabra, en la columna dos encontramos el resultado del coseno de los dos vectores (palabras o textos) que nos permite dar origen a la columna tres que es listado de las palabras o vecindario semántico, que proporcionan gran información de la palabra.

	LSA Similarity	Term
1	0.99	cat
2	0.84	cats
3	0.74	meow
4	0.68	begind
5	0.68	mouse
6	0.67	yellowest
7	0.67	outwait
8	0.65	tabby
9	0.65	braying
10	0.64	bray
11	0.63	purrs
12	0.62	slimmest
13	0.62	tojo
14	0.62	bojo
15	0.62	thrumm
16	0.62	siamese
17	0.61	donkey
18	0.60	birding
19	0.59	scrounging
20	0.59	tucker

Tabla 1. Muestra las correlaciones existentes entre la palabra "gato" y las 20 palabras más cercanas en el espacio semántico correspondiente.

[6] Venegas, René. Revista Signos. Documento LSI disponible en http://www.scielo.cl/scielo.php?pid=S0718-09342003005300008&script=sci_arttext consultado en marzo de 2007.

A partir del cálculo del coseno de cada vector, se puede establecer sobre que documentos conoce más LSI, gracias a una medida denominada longitud del vector, donde entre mas larga sea la longitud del mismo más conocimiento existe sobre la temática, esto genera que palabras como "el", "de", "los" entre otras se encuentren con vectores bajos de los que no casi no conoce.

El modelo ISL es de gran utilidad en la recuperación de información, ya que permite aprovechar de una mejor manera, las relaciones entre los términos de modo que se logre coherencia y comprensión textual. Evitando modelos como los buscadores de Internet, que limitan la búsqueda a la coincidencia de palabras dentro de un documento y arrojando resultados fuera de contexto.

Sin embargo la ISL no solo se maneja en el campo de la recuperación de información, también presta un enorme apoyo en la educación, por medio de los simuladores de ensayos, donde los estudiantes le solicitan a la maquina que lea y analice sus escrito, la maquina logra establecer, la coherencia semántica del documento y da las respectivas opiniones y sugerencias, con excelentes resultados según los mismos usuarios[7].

[7] Disponible en www.lsa.colorado.edu consultado en abril de 2007

BIBLIOGRAFÍA

- **JORGE DE Y BOTANA, Guillermo**. El Análisis de la Semántica Latente y su aportación a los estudios de usabilidad disponible en http://www.nosolousabilidad.com/articulos/analisis_semantica_latente.htm
- **JORGE DE Y BOTANA, Guillermo** Adecuación de ruta: nuevo índice basado en el Análisis de la Semántica Latente http://www.nosolousabilidad.com/articulos/adecuacion_ruta.htm
- **SEVILLANO DOMÍNGUEZ, Xavier; PUJOL Francesc Alýas y SOCOR´O CARRI´E,Joan Claudi.** Extracción de trópicos independientes para la clasificación de textos.
- **JORGE DE Y BOTANA, Guillermo; OLMOS Ricardo, LEÓN José A**
- **Manuscrito no definitivo. Análisis de la Semántica** Latente (LSA) y estimación automática de las intenciones del usuario en diálogos de telefonía (call routing). Universidad Complutense de Madrid; Departamento de Procesos Cognitivos, Facultad de Psicología (Universidad Autónoma de Madrid) disponible en
- **LANDAUER,Th.; Foltz,P. y LAHAM,D.** (1998) An Introduction to Latente Semantic Analysis. Discourse Processes.25(2&3), 259-284
- **VENEGAS V., René.** Análisis Semántico Latente: una panorámica de su desarrollo. Rev. signos. [online]. 2003, vol.36, no.53 [citado 16 Mayo 2007], p.121-138. Disponible en la World Wide Web: <http://www.scielo.cl/scielo.php?script=sci_arttext&pid=S0718-09342003005300008&lng=es&nrm=iso>. ISSN 0718-0934
- **KINTSCH, W**. (2002) On the notions of theme and topic in psychological process models of text comprehension. En Louwerse, M. y van Peer, W. (Eds.) Thematics, Interdisciplinary Studies. 157-170. Amsterdam: Benjamins
Muñoz Jiménez, Félix. Modelos alternativos en IR. Disponible en: http://www.lapresentacion.com/madrid/ModelosAlternativos_IR.htm consultado en marzo de 2007.