LSA (Latent Semantic Analysis) en registros médicos

Gabriel Slomovitz

gslomo@gmail.com

**Resumen**. Latent Semantic Analysis es un método novedoso para extraer los componentes principales de un cuerpo de texto que ha sido utilizado para categorización y búsqueda de información. Sin embargo, por los resultados sorprendentes en lo que refiere a similitud con procesos realizados por humanos se convierte en una teoría promisoria para explicar procesos cognitivos. En este trabajo se detalla el principio de funcionamiento del método, basado en la técnica algebraica conocida como descomposición en valores singulares y se analizan algunos estudios para su aplicación en registros médicos.

1. Introducción a LSA

LSA (Análisis Semántico Latente) es un método que permite analizar texto y extraer relaciones de similitud de significado entre palabras o entre pasajes de palabras. LSA utiliza algunas operaciones algebraicas para representar a las palabras en un espacio multidimensional. En particular, aplica una técnica de descomposición matricial denominada descomposición en valores singulares (SVD por sus siglas en inglés) para extraer los componentes principales de la representación del texto. Estas representaciones permiten simular algunos procesos cognitivos que ocurren en el cerebro, tales como: reconocimiento de vocabulario, comprensión de lenguaje, evaluación de composición de relato [1].

Las estimaciones producidas por LSA son inferidas a partir de operaciones matemáticas que logran extraer relaciones subyacentes que no son detectables utilizando métodos más directos como análisis de co-occurrencia , contigüidad, correlacionamiento de palabras u otros métodos estadísticos.

LSA utiliza como entrada solamente el texto a analizar, y si bien tiene limitaciones ya que no utiliza el orden de las palabras ni relaciones sintácticas o lógicas, consigue resultados que sorprenden por su similitud con análisis realizados por humanos.

1. Funcionamiento

LSA representa el texto a analizar en una matriz que se conforma del siguiente modo. Cada fila de la matriz corresponde a una palabra en particular, y cada columna corresponde al pasaje o contexto en el que aparece. Las celdas de la matriz contienen la cantidad de veces que la palabra aparece en el pasaje. Luego se aplica el método de descomposición en valores singulares (SVD) que permite descomponer una matriz rectangular ***X*** (de ***t*** columnas por ***p*** filas) en el producto de tres matrices:

***X = T \* S \* PT*** (1)

Donde ***X*** es la matriz original, ***T*** es una matriz de ***t*** x ***r*** de columnas ortonormales, ***P*** es una matriz ***p*** x ***r*** de columnas ortonormales y ***S*** es una matriz de ***r*** x ***r*** diagonal ordenada de manera decreciente. Los componentes de la matriz ***S*** son los valores singulares, y las matrices ***T*** y ***P*** corresponden a los vectores de palabras y de pasajes, respectivamente. Puede demostrarse que es posible descomponer cualquier matriz rectangular de esta manera [1].

Se puede reducir la dimensionalidad de la matriz original utilizando para su reconstrucción solo algunos coeficientes de la matriz diagonal ***S***,tal como se describe en (2), donde se utilizan los primeros ***k*** valores singulares.

***X’ = Tk \* Sk \* PkT*** (2)

Esto es equivalente a aproximar por mínimos cuadrados [2]; ***X’*** es la mejor aproximación a ***X*** por mínimos cuadrados. Es justamente en esta reducción de dimensiones que reside la gran capacidad del método para lograr extraer los componentes principales de los datos a analizar.

Se describe a continuación un ejemplo simple de aplicación del método [1], utilizando pasajes de texto que corresponden a nueve artículos técnicos sobre “Human Computer Interaction” y cuatro sobre “Graph theory”. La *Figura 1* muestra el texto original, los títulos de ambos temas aparecen agrupados en **c** y **m** respectivamente.



*Figura 1 - Texto original*

La matriz ***X*** se conforma según se muestra en *Figura 2.*



*Figura 2 - Matriz palabra-pasaje*

En la *Figura 3* se presenta el resultado de la descomposición en valores singulares (SVD) y en la *Figura 4* se muestra la reconstrucción de la matriz ***X*** utilizando solamente dos dimensiones, es decir utilizando solo las dos primeras columnas de las matrices.



*Figura 3 - Matrices correspondientes a la descomposición en SVD*



*Figura 4- Matriz X reconstruida*

En términos generales, la matriz reconstruida permite describir un segmento de texto a partir de la cuantificación de las palabras que lo componen. A su vez, cada palabra puede verse como la suma de sus aportes en cada segmento. A partir de estas cuantificaciones, pueden estimarse relaciones de similitud entre palabras o entre pasajes de texto. Esta idea se describe con más detalle a continuación comparando las filas correspondientes a las palabras *human*, *user* y *minors.* En la matriz original se observa que la palabra *human* nunca aparece en el mismo pasaje que *use*r y *minor.* El coeficiente de correlación Spearman [1] entre *human* y *user* es -0.38 y entre *human* y *minors* 0.29. Sin embargo, en la matriz reconstruida los coeficientes de correlación se transforman en 0.94 para *human-user* y -0.83 para *human-minor.* Pese a que *human* y *user* nunca aparecen en el mismo pasaje, el procedimiento de reducción de dimensiones permite descubrir relaciones subyacentes. Como los contextos en los que aparecen las palabras tienen un significado similar, LSA las representa como similares y lo opuesto ocurre para *human* y *minor*.

En lo que refiere a comparaciones de similitud de palabras o pasajes, la manera más simple es calcular el coseno entre los vectores que resultan de la descomposición en valores singulares y su posterior reducción de dimensiones. Adicionalmente, pueden implementarse otras medidas de similitud como distancia Euclideana o “city block distance” [2].

1. Aplicaciones en registros médicos

LSA ha sido utilizado inicialmente como método de búsqueda de información y categorización de texto. Sin embargo, su habilidad para descubrir estructuras escondidas y asociar objetos similares lo convierten en una técnica adecuada para modelar ciertos procesos que ocurren en el cerebro [5]. Esto abre un enorme campo de investigación en el ámbito de ciencias cognitivas. En lo que refiere a aplicaciones médicas en particular, LSA ha sido propuesto como método para análisis de historias clínicas y registros médicos en general. A continuación, se analizan algunos trabajos publicados en esta línea.

En [6] se plantea que los sistemas actuales de asistencia computacional para la toma de decisiones clínicas (CDS por sus siglas en inglés) dependen de bases de datos construidas con estructuras hechas a mano y en general aprovechan muy poco la riqueza de información que se presenta en los registros médicos. Los autores sostienen que un sistema CDS que utilice lenguaje natural, y permita realizar inferencias sin intervención humana es de gran interés en esta área. Los CDS actuales son extremadamente rígidos y restrictivos, mientras que la interacción doctor-paciente no lo es. Los síntomas rara vez son formulados de forma precisa y es habitual que no encajen en las estructuras de bases de datos médicas. En consecuencia, se propone el uso de LSA para medir la relación semántica entre síntomas y enfermedades. Dado que LSA no distingue algunas estructuras del lenguaje, tales como negaciones, muy importantes en el ámbito médico, los autores plantean una mejora al mecanismo tradicional de LSA. El texto es pre-procesado a los efectos de agregar una etiqueta a las palabras negadas. Por ejemplos si las palabras “sin fiebre” aparecen en el texto, son sustituidas por “fiebre\_negada”. De modo que cualquier palabra y su negación pasan a ser dos palabras distintas para LSA. Esto se consigue implementando un algoritmo que detecta en primera instancia condiciones clínicas y luego determina si éstas aparecen negadas o no.

Para el estudio realizado en [6] se genera un espacio semántico en base a más de 3000 documentos médicos sobre enfermedades (unos 250 MB) extraído de Healthline.com [7]. A los efectos de evaluar el método, se eligen combinaciones de enfermedades y síntomas tomadas aleatoriamente de “*Professional Guide to Signs and Symptoms*” [8]. Se crean 1000 muestra aleatorias como base para la comparación (ver ejemplo en Tabla 1). Luego de aplicar LSA, se mide la similitud del par enfermedad-síntoma real en base al coseno y se lo compara con el coseno de los pares enfermedad-síntoma aleatorios. Se utiliza un clasificador del tipo *Naive-Bayesian* para determinar el mejor umbral y poder así clasificar automáticamente pares reales y pares aleatorios.

Para evaluar los resultados se realiza un análisis de varianza (Anova) entre las medidas de similitud de los pares reales y aleatorios. Tal como se muestra en la Figura 5, los resultados reflejan una diferencia significativa, con altos valores de coseno para los pares reales y más bajos para los aleatorios. Esto prueba que LSA es capaz de relacionar con buen grado de precisión los síntomas con las enfermedades correspondientes. Este trabajo es una muestra de las capacidades del método y del gran potencial de técnicas de este tipo para inferir conclusiones automáticamente a partir de texto médico sin la necesidad de bases de datos estructuradas o intervenciones manuales.



*Tabla 1 - Enfermedad-Síntoma real y aleatorio*



*Figura 5 – Anova entre pares reales y aleatorios*

Una línea de investigación distinta es planteada en [9], donde se propone un sistema de búsqueda de información médica denominado “modelo de enriquecimiento semántico”. La idea es identificar el contexto de la búsqueda para luego comparar el o los términos del *query* con una matriz generada a partir de LSA. El proceso, representado en la Figura 6, consta de los siguientes pasos. En primer lugar, se realiza una consulta relacionada con el contexto elegido para el estudio: cáncer de cerebro, en la ontología del National Cancer Institute. Como resultado se tiene un conjunto de términos que representan las clases relacionadas con la búsqueda. Luego estos términos son comparados con la matriz semántica de similitud (MSS). Dicha matriz fue obtenida a partir de la aplicación de LSA a artículos de un repositorio médico sobre el tema en particular (cáncer de cerebro). Cada fila y cada columna de MSS corresponde a un término. De modo que cada celda MSSij indica el grado de similitud (entre 0 y 1) entre el término *i* y el término *j.* Para determinar las palabras con mayor grado de similitud con un término cualquiera *t*, se buscan los valores más altos dentro de la fila o la columna correspondiente a *t*. En el caso que *t* no figure en la matriz, se realiza la búsqueda en base a sinónimos obtenidos de WordNet. A los efectos de evaluar el sistema, los autores utilizaron como repositorio 1125 artículos de la revista Plos One publicada por Public Library of Science. Se comparan los resultados obtenidos de distintos *queries* entre el modelo propuesto y métodos existentes. Para medir la efectividad de la búsqueda se utilizan dos conceptos del ámbito *de Information Retrieval*: precisión y sensibilidad (o *recall*). La precisión se define como la cantidad de instancias recuperadas relevantes sobre la cantidad total de instancias recuperadas. Mientras que la sensibilidad corresponde a la cantidad de instancias relevantes recuperadas sobre la cantidad total de instancias relevantes. Los resultados se muestran en la Tabla 2. Como conclusión se establece que el trabajo demuestra la importancia de tomar en cuenta no sólo la sintaxis sino también la semántica de los términos de búsqueda en sistemas de recuperación de información.



*Figura 6 – Proceso de búsqueda*



*Tabla 2 – Resultados de precisión y sensibilidad*

Existen otros modelos de procesamiento de lenguaje natural basados en LSA que incorporan elementos estadísticos: pLSA (probabilistic LSA) y LDA (Latent Dirichlet Allocation). Estas teorías se basan en que los documentos pueden verse como un conjunto de temas distribuidos según alguna distribución de probabilidades (distribución de Dirichlet para LDA). Sin entrar en mayor detalle, en [10] y [11] se utilizan estas teorías para analizar registros médicos. Se muestra que estos sistemas pueden ser aplicados exitosamente en el análisis de historias clínicas, diagnóstico, prescripciones, así como en informes de imágenes médicas.

1. Conclusiones

Las operaciones algebraicas realizadas por LSA, en particular SVD, pueden verse como un método para identificar y ordenar las dimensiones sobre las cuales los datos exhiben mayor variación. Una vez identificadas las dimensiones, pueden aproximarse los datos por su proyección sobre este espacio. SVD permite reducir los datos a un espacio dimensional más pequeño que expone la subestructura de los datos originales y además los ordena de mayor a menor variación. LSA, simplemente ignora las variaciones por debajo de cierto umbral, reduciendo drásticamente la cantidad de datos, pero preservando los principales componentes.

Las mejoras en las últimas décadas de la capacidad de procesamiento computacional han abierto un enorme campo para la aplicación de LSA en diversos ámbitos. Los resultados obtenidos cuando se somete a LSA a ciertas tareas, sorprenden por su similitud con los resultados obtenidos cuando esas mismas tareas son realizadas por personas. Hay que tener en cuenta, que para obtener esos resultados las personas deben “entender” las palabras y pasajes de texto analizadas. Es por esto último que LSA es visto no solamente como un método para categorización o búsqueda de texto, sino como una teoría computacional para análisis del lenguaje y la comprensión.

En el ámbito de registros médicos, LSA puede ser utilizado para el análisis automático de historias clínicas, como sistema de recomendación para diagnósticos de enfermedades, como sistema automático de anotación de imágenes médicas, etc. Cabe señalar, que por las características de LSA, no es necesario generar estructuras sintácticas adicionales, ni codificaciones de ningún tipo, a diferencia de la mayoría de los sistemas de procesamiento de lenguaje natural. Es por ello que resulta un mecanismo ideal para su aplicación en lenguaje médico. En resumen, existe un vasto campo de trabajo futuro en la implementación de técnicas de minado de texto del estilo de LSA en distintos ámbitos médicos. Dado que existen diferentes proyectos de informatización en instituciones de salud en nuestro país, así como en AGESIC, sería de gran interés poder explorar la aplicación de LSA en registros médicos nacionales. De este modo se podrían generar primeras experiencias en el rubro a partir de datos reales que sirvan tanto para demostrar la capacidad de este tipo de tecnologías, así como desarrollar una nueva línea de investigación a nivel nacional.

1. Bibliografía

[1] T. K. Landauer, P. W. Foltz, D. Laham, ”Introduction to Latent Semantic Analysis,” Discourse Processes,25, pp .259-284, 1998.

[2] T. K. Landauer, S. Dumais (2008), Scholarpedia, 3(11):4356.<http://www.scholarpedia.org/article/Latent_semantic_analysis>

[3] T. K. Landauer, D. S. McNamara, S. Dennis, W. Kintsh, "Handbook of Latent Semantic Analysis,” 2011.

[4] K. Baker, “Singular Value Decomposition Tutorial,” 2005.

[5] J. C. Valle-Lisboa, E. Mizraji, “The uncovering of hidden structures by Latent Semantic Analysis,” Information Sciences, 177, pp. 4122–4147, 2007.

[6] V. Dada, King-Ip, M. Louwerse “Capturing disease-symptom relations using higher-order co-occurrence algorithms,” IEEE International Conference on Bioinformatics and Biomedicine Workshops (BIBMW), pp. 816-821, 2012.

[7] HealthLine, "Healthline," oct 2011. [Online]. Available: <http://www.healthline.com/>

[8] W. W. Lippincott, *Professional Guide to Signs and Symptoms,* ser. Professional Guide to Signs and Symptoms. LippincottWilliams & Wilkins, 2010.

[9] J.C. de Araujo, J.M. Parente, L. Garcia, “Recovery Medical Articles using Semantic Enrichment

Method,” 11th International Conference on Signal-Image Technology & Internet-Based Systems, pp. 705-711, 2015.

[10] Xinyu Jin; Wentao Ma; Yunze Li, “Medical Record Text Analysis Based on Latent Semantic Analysis,” 8th International Symposium on Computational Intelligence and Design (ISCID), Volume: 2, pp. 108-110, 2015.

[11] Bo Li, Ke Wang, “Computer aided diagnosis semantic model for the report of medical image via LDA and LSA”, IT in Medicine and Education (ITME), 2011 International Symposium on, Volume 1, pp. 699-703, 2011.