Motores de Búsqueda Web: Estado del arte en Probabilistic Latent Semantic Analysis y en Latent Dirichlet Allocation aplicado a problemas de acceso a la información en la Web.

José Alberto Benítez Andrades y Juan Antonio Valbuena López

Octubre 2011

En este trabajo se introducen los dos métodos de probabilidad estadística PLSA y LDA. En una primera parte, se introduce y se explica con detalle el PLSA, teniendo en cuenta algunas de las aplicaciones realizadas con este método en los ultimos años. En segundo lugar se describe la técnica LDA, detallando también algunas de sus distintas aplicaciones. Y como parte final del trabajo, se realiza una comparación entre los dos métodos, comentando sus diferencias, ventajas e inconvenientes en función del marco donde se analizan.

Índice

1. Introducción al PLSA.

1.1. Qué es "Probabilistic Latent Semantic Analysis" (PLSA).

El Análisis Probabilístico Semántica Latente (PLSA), también conocido como Indexación Semántica Latente Probabilística (PLSI, especialmente en los círculos de recuperación de información) es una técnica estadística para el análisis de dos modos y los datos de co-ocurrencia. PLSA ha evolucionado desde el análisis semántico latente (LSA), la adición de un modelo probabilístico lo hizo más sólido.

PLSA tiene aplicaciones en la recuperación de la información y filtrado, procesamiento del lenguaje natural, aprendizaje automático a partir del texto, y áreas relacionadas. Fue introducido en 1999 por **Jan Puzicha** y **Thomas Hofmann**, y se relaciona con la factorización de matrices no negativas.

En comparación con el análisis semántico latente estándar que se deriva de álgebra lineal y "downsizes" las tablas de ocurrencia (por lo general a través de una descomposición de valor singular), el análisis semántico latente probabilístico se basa en una mezcla de derivados de la descomposición de un modelo de clases latentes. Esto da como resultado un enfoque más de principios que tiene una sólida base en las estadísticas. Teniendo en cuenta las observaciones en forma de co-ocurrencias (w, d) de las palabras y documentos, en los modelos PLSA la probabilidad de cada co-ocurrencia se percibe como una mezcla de distribuciones multinomial condicionalmente independientes.

1.2. ¿ En qué consiste ?

Los modelos probabilistas parten de distribuciones de probabilidad para representar el conocimiento encerrado en el lenguaje. En general, estos modelos están diseñados para aplicaciones específicas, típicamente para clasificación de textos y recuperación de información, puesto que el conocimiento que son capaces de recoger es limitado.

El modelo PLSA fue presentado y aplicado por primera vez en la minería de texto por Hoffman. En contraste con el algoritmo estándard LSI, que utiliza la norma de Frobenius como un criterio de optimización, el modelo PLSA se basa en el principio de probabilidad máxima, que es derivado de teorías estadísticas dudosas. Básicamente, el modelo PLSA se basa en el modelo estadístico llamado modelo de aspecto, que puede ser utilizado para identificar relaciones semánticas ocultas mediante actividades de co-ocurrencia.

Teóricamente, podemos ver las sesiones de los usuarios sobre las páginas web como actividades co-ocurrentes en el contexto de la minería de uso web, para concluir el patrón de uso latente. Dado el modelo de aspecto sobre el patrón de acceso de usuario en el contexto de la minería web, primero se asume que hay un factor latente de espacio Z = (z1, z2, $\cdots$ zk), y cada lista de datos derivada de la observación de co-ocurrencia (s_i, p_j) (por ejemplo, la visita de la página p_j en la sesión de usuario s_i) es asociada con el factor z_k. Acorde a este punto de vista, el

modelo de aspecto puede concluir en que existen diferentes relaciones entre los usuarios web o las páginas correspondientes a diferentes factores. Además, los diferentes factores pueden ser considerados para representar los correspondientes patrones de acceso a usuario. Por ejemplo, durante un proceso de minería de uso web en una tienda online, nosotros podemos definir que existen k factores latentes asociados con k tipos de patrones de navegación, como . Furthermore, the different factors can be considered to represent the corresponding user access pattern. For example, during a Web usage mining process on an e-commerce website, we can define that there exist k latent factors associated with k kinds of navigational behavior patterns, así como el factor z_1 para los que tienen interest en los productos deportivos, z_2 para productos con interés en venta and y z_3 para buscar a través de la variedad de páginas de productos en diferentes categorías.

De esta manera, cada uno de los datos de observación de co-ocurrencia (s_i, p_j) puede transmitir interés de los usuarios de navegación mediante la asignación de los datos de observación en el espacio latente k-dimensional de los factores. El grado, a la que este tipo de relaciones se "explican" por cada uno de los factores, se obtiene una distribución de probabilidad condicional asociada con los datos de uso de la Web. Por lo tanto, el objetivo de emplear el modelo PLSA, es determinar la distribución de probabilidad condicional, a su vez, para revelar las relaciones intrínsecas entre los usuarios de la Web o las páginas basadas en un enfoque de inferencia de probabilidad. En una palabra, el modelo PLSA es modelo y el comportamiento de navegación del usuario puede concluir en un espacio semántico latente, e identificar el factor latente asociado. Antes de proponer el algoritmo basado PLSA para la minería uso de la Web, es necesario introducir la formación matemática del modelo PLSA, y el algoritmo que se utiliza para estimar la distribución de probabilidad condicional.

- $\mathrm{P}(s_i)$ representa la probabilidad que será observada en una sesión de usuario particular s_i,
- $\mathrm{P}(z_k \mid s_i)$ representa una probabilidad de una sesión de usuario específica sobre la clase z_k factor latente,
- $\mathrm{P}(p_j \mid z_k)$ representa la probabilidad de distribución de la clase condicional de las páginas sobre una variable latente z_k.

Basándonos en las siguientes definiciones, el modelo PLSA puede expresarse de la siguiente forma:

- Seleccionando una sesión de usuario s_i con una probabilidad de $\mathrm{P}(s_i)$,
- Escogiendo un factor oculto $z_k$con una probabilidad $\mathrm{P}(z_q \mid s_i)$,
- Generando una página $p_j$con una probabilidad $\mathrm{P}(p_j \mid z_k)$;

Y como resultado, obtendríamos una probabilidad ocurrente de un par observado $(z_q \mid p_j)$ que adopta el factor variable z_k. Traduciendo este proceso en un modelo de probabilidad de resultados en la expresión:

P(s_i | p_j) = P(s_i) · P(p_j |s_i)
donde P(p_j ,s_i) = $\sum_{z\epsilon Z}$P(p_j | z) · P(z |s_i)
Aplicando la fómrula Bayesiana, una versión reparametrizada puede ser transformada en la ecuación siguiente:
P(p_j,s_i) = $\sum_{z\epsilon Z}$P(z) P(s_i | z) P(p_j | z)
Siguiendo el principio de probabilidad, nosotros podemos determinar la probabilidad total de la observación como:

$$Li = \sum_{s_i \in S, p_j \in P} m(s_i, p_j) \cdot \log P(s_i, p_j)$$

donde $m(s_i, p_j)$ corresponde a la entrada de la matriz asociada de la sesión de páginas vistas con la sesión s_i y la página vista p_j. Para maximizar la probabilidad total, es necesario generar repetidamente las probabilidades condicionales de P(z) , P(s_i | z) y P(p_j | z) utilizando el uso de los datos de observación. Sabiendo las estadísticas, el algoritmo de EM (Expectation-Maximization) es un procedimiento eficiente para mejorar la estimación de probabilidad máxima en el modelo latente variable. Generalmente se necesitan dos pasos para implementar en el procedimiento alternativamente: (1) el paso de "Expectation" (E), donde las probabilidades posteriores son calculadas por los factores latentes basados en las estimaciones actuales de la probabilidad condicional, y la Maximización (Maximization (M)), que es el paso donde las probabilidades estimadas condicionales se actualizan e intentan maximizar la probabilidad basadas en las probabilidades posteriores computadas en el paso anterior.

Todo el procedimiento se da como sigue: En primer lugar, dados los valores al azar inicial de P (z), P (si | z), P (pj | z), entonces, en el paso E, podemos simplemente aplicar la fórmula bayesiana para generar el siguiente variable basada en la observación de su uso:

$$P(z_k|s_i, p_j) = \frac{P(z_k)P(s_i|z_k)P(p_j|z_k)}{\sum_{z_k \in Z} P(z_k)P(s_i|z_k)P(p_j|z_k)}$$

además, en el paso de Maximización, se computa:

$$P(p_j|z_k) = \frac{\sum_{s_i \in S} m(s_i, p_j)P(z_k|s_i, p_j)}{\sum_{s_i \in S, p_j' \in P} m(s_i, p_j')P(z_k|s_i, p_j')}$$

$$P(s_i|z_k) = \frac{\sum_{p_{j\in} \in P} m(s_i, p_j)P(z_k|s_i, p_j)}{\sum_{s_i' \in S, p_j \in P} m(s_i', p_j)P(z_k|s_i', p_j)}$$

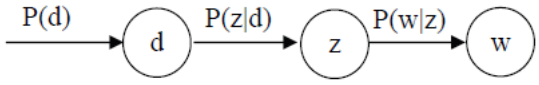
$$P(z_k) = \frac{1}{R} \sum_{s_i \in S, p_j \in P} m(s_i, p_j)P(z_k|s_i, p_j)$$

donde $$R = \sum_{s_i \in S, p_j \in P} m(s_i, p_j)$$

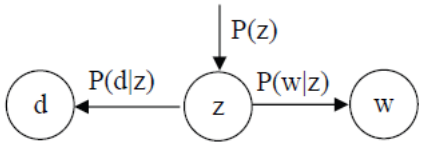
La implementación iterativa de la E-paso y paso-M se repite hasta que Li está convergiendo hacia un límite de locales óptimos, lo que significa que los resultados calculados pueden representar las estimaciones de probabilidad óptima de los datos de uso de la observación. De la formulación anterior, es fácil de encontrar que la complejidad computacional del modelo PLSA es O (mnk), donde m, n y k denotan el número de sesiones de usuario, páginas Web y los factores latentes, respectivamente.

Por ahora, hemos obtenido la distribución de probabilidad condicional de P (ZK), P (si | zk) y P (pj | zk) mediante la realización de la E y el paso M iterativa. La distribución de probabilidad que se estima que es correspondiente a la máxima verosimilitud local contiene la información útil para inferir el uso de factores semánticos, el usuario performing Web sesiones de la agrupación que se describen en las secciones siguientes.

El modelo de Análisis de Semántica Latente Probabilista PLSA (Probabilistic Latent Semantic Analysis) también podemos considerar que propone un modelo de probabilidad, dados un documento y una palabra, de dos maneras distintas, como muestra la siguiente figura (es otra forma de representarlo) . En el caso a) el modelo es asimétrico y en el caso b) simétrico, siendo d el documento, q la palabra, z la categoría o tópico y P la función de probabilidad.



a)



b)

Al igual que en LSA, el experto ha de definir clases, tópicos o pasajes en los que incluir los textos y palabras. Así pues, los modelos de la figura responden a las expresiones siguientes:

a) P(d,w) = $\sum_z P(w|z) \cdot P(z|d)$

b) P(d,w) = $\sum_z P(z) \cdot P(d|z) \cdot P(w|z)$

Estos modelos son posteriormente ajustados mediante un algoritmo EM (*Expectation Maximization*), obteniendo así tres matrices: U, que contiene las probabilidades de los documentos dadas las categorías, V, que contiene las probabilidades de las palabras dadas las categorías y la matriz diagonal $\sum$, que contiene las probabilidades de las categorías. El modelo queda definido finalmente como $P = U \bullet V \bullet \sum$.

Aunque se pueda establecer cierta analogía con la descomposición en valores singulares de LSA, la matriz P presenta ciertas diferencias ventajosas:

- P es una distribución de probabilidad bien definida y los factores tienen un claro significado probabilista, contrariamente a LSA.
- Las direcciones de los vectores en el espacio LSA no tienen interpretación. En PLSA son distribuciones multinomiales de palabras.
- La elección del número de dimensiones del espacio tiene un trasfondo teórico en PLSA. En LSA se hace de manera experimental.

Experimentos realizados en recuperación de información demuestran que PLSA obtiene una precisión entrono al 10 % mejor que LSA en varios corpus de textos.

El modelo de Localización de Dirichlet Latente LDA proporciona, según sus autores, una semántica completa probabilística generativa para documentos. Los documentos se modelan mediante una variable aleatoria oculta de Dirichlet. Al igual que en PLSA, asume un conjunto de categorías predefinidas. Cada categoría se representa como una distribución multinomial sobre el conjunto de palabras del vocabulario. El modelo queda descrito por la siguiente expresión:

$$P(d) = \int \vartheta \; [\; \Pi n \; \Sigma z \; P(wn|zn;\beta) \cdot P(zn|\vartheta) \;] \cdot P(\vartheta;\alpha) \; \delta\vartheta$$

siendo $P(\vartheta;\alpha)$ una distribución de Dirichlet, $P(zn|\vartheta)$ una multinomial que indica el grado en que el tema zn es tratado en un documento y β una matriz de clases por palabras del vocabulario. De esta manera, la probabilidad de un documento depende de las probabilidades de que sus palabras denoten ciertas categorías dentro de una distribución de Dirichlet. Para aprender e inferir en el modelo usan un algoritmo EM análogo al modelo PLSA descrito anteriormente.

El modelo de Mezcla de Unigramas [Nigam et al., 2000] es un modelo sencillo y muy similar a los dos sistemas anteriores. Está descrito por la siguiente expresión:

$$P(d) = \Sigma z \; (\; \Pi n \; P(wn|z) \;) \cdot P(z)$$

Como se aprecia, la probabilidad de un documento depende de las probabilidades de que sus palabras pertenezcan a las categorías.

Experimentos en clasificación de textos y en recuperación de información muestran la superioridad del modelo LDA con respecto a PLSA y al modelo de Mezcla de Unigramas. Además, LDA recoge la posibilidad de que un documento contenga más de una categoría temática, al contrario que la Mezcla de Unigramas, y no está condicionado por los ejemplos de entrenamiento, como es el caso de PLSA.

1.3. Aplicaciones del PLSA en distintos estudios.

Este método probabilístico latente, desde su creación en 1999 por Hoffman, ha sido utilizado con diferentes fines y modificado por disintos investigadores, con la intención de mejorar su funcionamiento aplicando distintos criterios.

Leyendo algunos de los proceedings de los últimos años, he seleccionado cinco estudios que me han parecido bastante interesantes, en los que se aplica la

técnica de PLSA, con distintas modificaciones para poder conseguir una mejora en los resultados de los experimentos realizados.

Los proceedings que he elegido son los siguientes:

- Adaptive Label-Driven Scaling for Latent Semantic Indexing. SIGIR 2010
- Topic-bridged PLSA for Cross-Domaint Text Classification. SIGIR 2010
- ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews. SIGIR 2011.
- Clickthrough-Based Latent Semantic Models for Web Search. SIGIR 2011.
- Regularized Latent Semantic Indexing. SIGIR 2011.

1.3.1. Adaptive Label-Driven Scaling for Latent Semantic Indexing.

En primer lugar, comentar que los autores de este proceeding son *Xiaojun Quan, Enhong Chen, Quiming Luo y Hui Xiong.* Pertenecen al departamento de ciencias de la computación de la Universidad de Ciencia y Tecnología de China (USTC).

Este trabajo de investigación, se trata principalmente de mejorar la técnica de LSI mediante la explotación de etiquetas de categoría. Específicamente, en la matriz de términos de documento, el vector para cada término apareciendo en etiquetas o semánticamente cercano a las etiquetas, se escala antes de realizar la técnica de SVD (Singular Value Descomposition [Descomposición de valor singular]) para aumentar su impacto en la generación de vectores singulares. Como resultado, las similitudes entre los documentos pertenecientes a una misma categoría, se incrementan. Además, se diseña una estrategia de escalado adaptativo para mejorar la utilización de estructuras de herencia para las categorías. Los resultados de este experimento muestran que el enfoque que proponen sus autores, mejora significativamente la actuación de la categorización de texto por herencia.

En su proceeding, destacan en la **introducción** que la Indexación Semántica Latente (LSI), es una técnica de recuperación y categorización de texto que cuenta con diferentes frameworks para poder aplicarlo y que ha recibido anteriormente distintas mejoras en otros estudios realizados por otros investigadores. Ellos proponen un enfoque de aplicación del LSI explotando las etiquetas de categoría, como indiqué anteriormente.

En la **metodología**, explican que en la categorización de texto, los términos en un documento que también aparecen en etiquetas de categorías son más efectivos categorizando el documento que otros términos. Se encargaron de estudiar una estrategia para impulsar el impacto. Su propuesta fue escalar los vectores de términos de etiquetas de categoría en la matriz de términos del documentos antes de impulsar el SVD.

Extienden el hecho de escalar una serie de términos que son similares a las etiquetas. Estos términos aparecen en las etiqeutas o son similares a las etiquetas

de categorías. Estos términos se llaman "label-relevant" (etiquetas relevantes). Estos términos se basan en la siguiente fórmula:

$$\text{label-relevant (t)} = \{s|\text{rank}(\text{sim}(s,t)) \leq l\}$$

En esta fórmula sim(s,t) representa la similitud entre s y t. Demuestran que mediante su método "label-driven scaling", se incrementa la similitud de una consulta con un documento de la misma categoría. Desarrollan su estudio de forma matemática y sacan una serie de conclusiones.

Explican, que en uno de los ejemplos, cuando la consulta y el documento pertenecen a la misma categoría, ellos tienen más probabilidades de tener un término "label-relevant".

Y en relación a la categorización de texto por herencia, explican que la organización de las categorías es mediante herencia, y que las que se encuentran en el nivel más inferior del árbol de herencia, son las más especificas.

Los experimentos o pruebas que realizan, consisten en colecciones de datos que tienen ya las categorías organizadas como taxonomías y cuya etiqueta para cada categoría está predefinida. Estos documentos son preprocesados siempre. Después de eliminar una "stopword", ellos se encargan de filtrar términos con menos de dos caracteres. Para decidir qué términos son "label-relevant", ellos utilizan LSI con la reducción de dimensión en 50. En el proceso de clasificación, el número de vecinos cercanos, se configura en 20. Finalmente, ellos comparan la actuación de dos variantes de su enfoque con dos enfoces: clasificadores kNN cuyas similitudes se obtienen en el espacio LSI; y clasificadores SVM de herencia, usando un núcleo lineal y unos parámetros con valores por defecto. La diferencia entre NADP y SLSI es que el formador aplica un escalado uniforme a todos los nodos en la herencia mientras el último aplica un escalado de adaptación. En muchos casos, para el escalado "label-driven" y para el escalado de adaptación, se mejora la clasificación de la actuación.

Como conclusiones, declaran que sus dos enfoques del LSI: escalado "label-driven" y de adaptación, mejoran los resultados en datos del mundo real, gracias a la categorización con herencia.

1.3.2. Topic-bridged PLSA for Cross-Domain Text Classification.

Este trabajo de investigación fue realizado por *Gui-Rong Xue, Wenyuan Dai, Qiang Yang y Yong Yu*. Pertenecen a la *Universidad de Ciencia y Tecnología Clearwater Bay, Knowloon, Hong Kong.*

En muchas aplicaciones web, como por ejemplo la clasificación de blogs, clasificación de grupos de noticias, o datos etiquetados, son escasos. Obtener etiquetas de un nuevo dominio, suele ser caro y consume mucho tiempo, mientras que puede haber un conjunto de datos etiquetado en un dominio distinto pero que se relaciona con el nuevo. Los métodos de clasificación de texto antiguos no permiten aprender cruzando distintos dominios. Estos investigadores proponen un algoritmo de clasificación de texto mediante el cruce de dominios que en definitiva, extiende el PLSA para integrar datos etiquetados y datos no

etiquetados que vienen de dominios distintos pero que están relacionados, en un modelo probabilístico unificado. A este algoritmo nuevo lo llaman Topic-bridged PLSA (TPLSA). El algoritmo consiste en explotar los temas entre dos dominios y transferir la base de conocimiento entre esos dominios mediante un "puente de temas" (topic-bridge), que ayuda a la clasificación de texto en el dominio de destino. Una ventaja única que tiene su método es la capacidad para extraer al máximo el conocimiento que luego puede ser transferido entre los dominios. Esto hace que este algoritmo sea de los mejores en cuanto a clasificación de texto se refiere.

Explican primeramente cuáles son las tareas que realizan los framework en el aprendizaje tradicional. Dan mucha importancia y recalcan varias veces que etiquetar nuevos dominios es costoso en cuanto a tiempo sobre todo. Hay que tener en cuenta también, que en una web que se actualiza con mucha frecuencia, es complejo tener todas las etiquetas actualizadas, si no se consigue de manera automática.

Lo que ellos proponen parte de dos conjuntos de datos D_L y D_U, que están relacionados pero pertenecen a distintos dominios. D_L representa el conjunto de datos etiquetados del dominio antiguo y D_U pertence al nuevo dominio y necesita ser clasificado. Las etiquetas que pertenecen a D_L y las que van a predecirse para D_U son creadas desde el mismo conjunto de etiquetas C. El objetivo es clasificar al completo el conjunto Du a través del dominio antiguo y su conjunto de datos D_L.

La principal ventaja de este algoritmo es que extendiendo el modelo PLSA para datos desde distintos dominios (los de entrenamiento y los de pruebas), permiten indicar partes de la base de conocimiento a través del TPLSA que son constantes entre diferentes dominios y partes que son específicas de cada uno. Esto permite transferir la base de conocimiento aprendida incluso cuando los dominios son diferentes.

En este proceeding, explican que existen distintos tipos de clasificadores de texto tradicionales, con aprendizaje supervisado y semi-supervisado, pero que no sirven para el experimento que desean hacer funcionar, la transferencia entre dominios distintos.

TPLSA: Definición del problema, aprendizaje y pruebas.

Los elementos principales para aplicar este algoritmo son: documento d que representa la instancia entrenada, que asu veces es asignada a una etiqueta única desde un conjunto temático $C = \{c1, \ldots, ck\}$. Un vocabulario de palabras $W = \{w1,\ldots,wv\}$ que es dado y representa una bolsa de palabras. A partir de estos elementos, se trabaja con los conjuntos de documentos etiquetados y sin etiquetar D_L y D_U.

Este modelo TPLSA se puede dividir en dos partes. En relación con el conjunto de documentos etiquetados del dominio antiguo, podemos decir que PLSA actúa en D_L x W de la siguiente manera:

$Pr(d_l|w) = \sum Pr(d_l|z)Pr(z|w)$

donde $d_l \epsilon D_L$es el documento del conjunto entrenado.

Para el conjunto de datos de prueba D_U acorde con la observación de D_U y W, podemos decir que D_U x W se define de la siguiente forma:

$Pr(d_u|w) = \sum Pr(d_u|z)Pr(z|w)$
donde $d_u \epsilon D_U$es el documento del conjunto de pruebas.

Teniendo en cuenta esto, mediante una serie de ecuaciones, se relacionan las probabilidades condicionales de los documentos de prueba y los documentos ya entrenados y se van creando las etiquetas para los documentos del nuevo dominio.

Una vez hecho esto, se optimiza mediante otra serie de ecuaciones de probabilidad y lógica y se aplica el algoritmo de EM para asegurar que el valor de las funciones cumplen la optimización.

Las pruebas que realizan parten de la base de 3 conjuntos de datos.

Conclusiones finales del TPLSA.

Después de realizar la evaluación de 11 conjuntos de datos, los resultados que obtienen estos investigadores son muy positivos, ya que muestran que el algoritmo propuesto logra mejorar la actuación con respecto a otros algoritmos de clasificación.

En un futuro, considerarán otros métodos de aprendizaje para adquirir los parámetros usados en el modelo TPLSA y considerar otroas tareas de clasificación relacionadas como la clasificación "multi-clase".

En definitiva, la mejora que le añaden al modelo PLSA natural, es la inclusión de dos conjuntos de datos de distintos dominios que pueden transferirse entre ellos las categorías de sus términos.

1.3.3. ILDA: Interdependent LDA Model for Learning Latent Aspects and their Ratings from Online Product Reviews.

Esta investigación fue realizada por *Samaneh Moghaddam y Martin Ester*, ambas pertenecen a la *School of Computing Science, Simon Fraser University, Burnaby, BC, Canada.*

Hoy en día hay muchísimos análisis y críticas de distintos productos en Internet, de hecho, existen foros especializados para analizar distintos tipos de productos, blogs y grupos de discusión. Un ejemplo de ello puede ser Xataka, que analiza los distintos teléfonos móviles que salen al mercado.

Sin embargo, para una persona que desea comprar un producto, es bastante complejo en muchas ocasiones, poder llegar a contrastar todos los análisis que hay sobre un mismo producto en distintas webs. Esta dificultad ha provocado que la minería de opinión haya creado una nueva rama de investigación.

Los análisis de productos se componen de distintos elementos: por ejemplo, el aspecto es un atributo de un producto, en el caso de una cámara, la pantalla es un atributo de la misma. Teniendo en cuenta distintos elementos, los usuarios suelen puntuar los distintos componentes del producto y al finalizar, se hace una media de todas las puntuaciones y esa es la puntuación que obtiene el producto.

Este equipo de investigación propone tres modelos probabilísticos gráficos que extraen los distintos aspectos y puntuaciones de los productos de análisis que hay en la red. Los primeros dos modelos extienden el standard PLSI y LDA para generar un resumen de puntuación de aspecto. Introducen el modelo ILDA (Interdependent Latent Dirichlet Allocation). Las pruebas que realizan en sus

experimentos, utilizan conjuntos de datos reales, obtenidos de Epinions.com y demuestran la mejora en la efectividad del modelo ILDA.

Trabajos relacionados y definición del problema.

En este proceeding se explica que los trabajos realizados hasta el momento relacionados con la minería de opinión, en su gran mayoría, han sido enfocados a la tarea de identificación, y han ignorado el problema de la predicción de puntuación. Aún así, existen trabajos enfocados a este problema, que intentan solucionarlo basándose en bolsas de palabras, recuperando distintas líneas de los análisis que hay en Internet.

Lo que proponen estos es utilizar dos modelos de asociación latente semántica, el primero agrupa palabras en un conjunto de aspectos acorde con el contexto y el segundo agrupa palabras acorde a sus estructuras latentes semánticas y al contexto de los disintos análisis que hay de cada producto.

Otros investigadores, presentan modelos probablísticos gráficos para el modelado de contenido de páginas independientes y de capas de información de páginas dependientes. Pero también ignoran las puntuaciones de estos análisis.

En definitiva, todas las ramas de investigación que emergían de este problema, iban enfocadas al mismo problema, pero no resolvían el problema que en este caso manejan.

Los investigadores de este proyecto, parten de un conjunto P = {P1,P2, ... , Pl} que representa un conjunto de los productos que pueden ser de categorías distintas. Para cada producto Pi hay un conjunto de análisis del mismo Ri = {d1, d2,...,dN}. Cada análisis dj consiste en un conjunto de frases de opinión como por ejemplo "gran zoom", "excelente calidad", etc. Se puede decir que el problema se descompone en los siguientes elementos:

- Aspecto: El aspecto es un atributo o componente de un producto que ha sido comentado en un análisis. Por ejemplo, "duración de la batería", en la frase "La duración de la batería de esta cámara es bastante corta".
- Puntuación: La puntuación es la satisfacción de un usuario con términos numéricos. La mayoría de las webs disponen de un sistema que va desde 1 hasta 5.
- Frase de opinión: Una frase de opinión f = < t, s > es un par de términos t y sentimientos s. Normalmente el término es un aspecto y el sentimiento expresa la opinión, por ejemplo: < duración de la batería, corta >.
- Análisis: El análisis es una bolsa de frases de opinión.
- Definición del problema: Dado un conjunto de análisis para un producto P, la tarea es identificar los k principales aspectos de P y entonces predecir la puntuación de cada aspecto.

Teniendo en cuenta esto, podemos tener la valoración de distintos aspectos de un teléfono móvil, en cual, se evalúan por ejemplo la cámara, las dimensiones, la calidad de grabación de vídeo, el sistema operativo y la velocidad del procesador. Teniendo en cuenta estos valores, se hace una media total de cada móvil y sale

una puntuación final. Pero se puede dar el caso, lógicamente, de que un móvil A tenga mayor puntuación media que un móvil B, pero si lo que buscamos es que tenga una cámara con mayor calidad, es posible que la B posea esta característica.

El objetivo de este proyecto de investigación es encontrar o predecir las distintas puntuaciones de estos "aspectos" analizando los distintos análisis de un mismo producto que se encuentran por internet, de forma automática.

Valor del PLSI en esta investigación.

Como comenté anteriormente, estos investigadores realizan pruebas utilizando tres modelos probabilísticos distintos. Yo voy a resumir únicamente la parte en la que se habla del PLSI.

El PLSI ha sido aplicado a distintos problemas de mineria de texto recientemente. Sin embargo, usaban solo esto para la identificación de aspectos y estos modelos no generaban una puntuación de los productos. Para ello, en este estudio, extienden el modelo del PLSI para identificar aspectos y predecir las puntuaciones, de forma simultanea. Siguiendo el modelo estándard gráfico, los nodos representan variables aleatorias y ejes indicando posibles dependencias. Los nodos "con sombra" son variables aleatorias observadas y los nodos "sin sombra" son variables aleatorias latentes. La parte exterior representa análisis o criticas y la parte interior representa opiniones. N y M son el número de críticas o análisis de productos y el número de opiniones en cada crítica, respectivamente. Si M es independiente de todas las otras variables de datos generadas (a y r), esto se ignora.

Para extender el PLSI en esta investigación, añadieron una segunda fila. Para cada producto P, se genera un modelo de PLSI para asociar un aspecto no observado am y una puntuación rm con cada observación, por ejemplo, cada frase de opinión f = $< t_m, s_m >$ en una crítica $d \epsilon R$. Se puede definir el modelo PLSI adaptado generativo de la siguiente forma:

1. Seleccionar una crítica d de R con probabilidad *P(d)*.
2. Para cada frase de opinión $< t_m, s_m >$, $m \epsilon \{1, 2, ..., M\}$
   a) Ejemplo $a_m \sim P(a_m |\ d)$ y $r_m \sim P(r_m |\ d)$.
   b) Ejemplo $t_m \sim P(t_m |\ a_m)$ y $P(s_m |\ r_m)$.

Traduciendo este proceso en una distribución de probabilidad, la expresión obtenida es la siguiente:

$$P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{t}, \boldsymbol{s}, \theta | \alpha, \beta_1, \beta_2) = P(\theta | \alpha)$$
$$\prod_{m=1}^{M} [P(a_m|\theta) P(r_m|\theta) P(t_m|a_m, \beta_1) P(s_m|r_m, \beta_2)]$$

Realizando una serie de modificaciones, la fórmula final para la utilización de este modelo es la siguiente:

$$P(\boldsymbol{a}, \boldsymbol{r}, \theta | \boldsymbol{s}, \boldsymbol{t}, \alpha, \beta_1, \beta_2) = \frac{P(\boldsymbol{a}, \boldsymbol{r}, \boldsymbol{t}, \boldsymbol{s}, \theta | \alpha, \beta_1, \beta_2)}{P(\boldsymbol{t}, \boldsymbol{s} | \alpha, \beta_1, \beta_2)}$$

Conclusiones de la investigación

Resumir los aspectos evaluados da una información muy útil a los usuarios que van a realizar uan compra. La propuesta de estos investigadores es un modelo que enseña un conjunto de aspectos de productos y sus correspondientes puntuaciones de una colección de críticas del producto que han sido preprocesadas en una colección de frases de opinión. Reconocen cuáles son los aspectos más importantes mediante su modelo ILDA y así realizar una evaluación más coherente.

Una de las desventajas de los modelos no supervisados es que la correspondencia entre agrupaciones generadas y variables latentes no son explícitas. Planean extender su trabajo, planean investigar la corresponencia entre las agrupaciones identificadas y los aspectos reales o puntuaciones.

1.3.4. Clickthrough-Based Latent Semantic Models for Web Search.

Este trabajo de investigación que trata sobre modelos semánticos latentes basados en clisk para las búsquedas web, ha sido realizado por *Jianfeng Gao, Kristina Toutanova y Wen-tau Yih*, pertenecientes al grupo de *Microsoft Research, One Microsoft Way Redmond, WA 98052 USA*.

Estos investigadores presentan dos modelos de "rankeo" de documentos para las búsquedas web que se basan en métodos de representación semántica y el enfoque estadístico para la recuperación de información. Asumiendo que una consulta es paralela a los títulos de los documentos clicados durante la consulta, se construyen a través de datos por click grandes cantidades de pares títulos-consulta; se enseñan dos modelos semánticos para estos datos. Uno es un modelo temático bilingüe dentro del framework del modelado de lenguaje. Esto rankea documentos para una consulta por la probabilidad de una consulta siendo una traducción basada en la semántica de los documentos. La representación semántica es independiente del idioma y del par título-consulta aprendido, con la suposición de que una consulta y sus pares de títulos comparten la misma distribución sobre los temas semánticos. El otro es un modelo de proyección discriminativo dentro de un framework de modelado de espacio vectorial. A diferencia del LSA y sus variantes, la matriz de proyección en el modelo que estos investigadores proponen, que es usado para mapear desde vectores de términos en un espacio semántico, aprenden discriminativamente que la distancia entre una consulta y su título, ambos representados como vectores en el espacio semántico proyectado, es más pequeño que entre la consulta y los títulos de otros documentos que no tienen clicks en esa consulta. Estos modelos son evaluados en la tarea de búsqueda Web usando un conjunto de datos reales. Los resultados que se muestran en esta investigación son altamente favorables.

Introducción.

Los motores de búsqueda modernos recuperan documentos buscando términos de forma literal dentro de esos documentos. Sin embargo, los métodos de búsqueda léxica pueden tener discrepancias según el lenguaje, por ejemplo, un concepto se expresa usando diferentes palabras y vocabulario en los documentos y en las búsquedas.

En las últimas décadas para poder crear buscadores semánticos, se han utilizado modelos de probabilidad semántica latente, creando bolsas de palabras en los cuales se agrupaban que significaban lo mismo, aunque fueran léxicamente diferentes.

El objetivo de esta investigación es desarollar nuevos modelos de ranking para la búsuqeda web, combinando, en un principio, los métodos de representación semántica y traducción estadística. En esta investigación se propone que la traducción entre una consulta y un documento, puede ser modelada de forma más efectiva mediante un mapeo de los mismo en representaciones semánticas que son independientes del lenguaje.

Como bien comenté en el resumen anterior, tienen 2 ramas de investigación, una basada en los clicks y otra en un modelo de proyección mediante vectores en el espacio.

Otros trabajos

Para poder realizar esta investigación, en primer lugar explican otros enfoques para poder enlazar las diferencias léxicas entre las consultas y los documentos para la recuperaión de información.

Presentan modelos de trducción estadistica: en cada documento se puntúa por probabilidad de traducción en una consulta. Se asumen una serie de documentos y consultas dentro de una bolsa de palabras, en los que se computa una probabilidad de aparición de palabras. Mediante esta probabilidad, consiguen verificar y crear las relaciones entre palabras léxicamente distintas, los documentos y sus consultas. A diferencia de los modelos LSA, que no mapean diferentes términos en grupos latentes semánticos, pero aprenden las relaciones de traducción de forma directa entre un término y un documento.

Por otro lado, los modelos de temas generativos, como por ejemplo el PLSA, fueron de los primeros utilizados en tareas de recuperación de información. Este modelo puede ser incorporado en un framework de modelado de lenguaje, bajo el cual los documentos son rankeados por sus probabilidades de generar una consulta. En PLSA, una consulta, vista como un documento corto, es generada por un documento utilizando el siguiente proceso: Primero, una distribución multinomial θde T temas para cada documento es seleccionada como la mejor distribución temática para el documento; en segundo lugar, un tema o categoría z es seleccionado para cada término con cierta probabilidad. Finalmente un término de consulta q se genera con una probabilidad, explicada en este trabajo anteriormente.

Y por último, los modelos de proyección lineal. El LSA es un ejemplo de este tipo de modelos. Similar a los modelos de tópicos, LSA también puede ser extendido para crear pares de tuplas de documentos comparables o paralelos.

Modelo de tema bilingüe.

Este modelo se puede ver como un caso especial de PLTM donde las consultas

de búsuqeda y los documentos web son asumidos para ser escritos en dos idiomas distintos y las conclusiones MAP son utilizadas en lugar de las conclusiones Bayesianas..

Para realizar la estimación MAP utiliza algoritmos de estimación EM, al igual que en el PLSA explicado en los primeros puntos de este trabajo.

Una vez realizada la estimación, se realiza la regularización posterior. Una consulta y un título, si son un par, se entiende que contienen palabras comunes relacionadas mediante las bolsas de palabras.

Después de realizar la regularización, teniendo en cuenta los distintos vectores y las distribuciones, se produce la clasificación o rankeo de documentos.

Conclusiones finales:

Según los resultados de los experimentos, se observa que usando PLSA sin combinar con ningún otro modelo, provoca una "herida" en la clasificación o ranking de documentos. Pero combinando de forma lineal PLSA y el modelo original de documento, mejora significativamente los resultados.

En un futuro intentarán explorar estrategias alternativas de combinación de modelos latentes semánticos y modelos de traducción para la recuperación de información. Por ejemplo, pueden formar un corpus títulos-consulta, donde ambos, consultas y títulos, sean etiquetados por temas o conceptos. Entonces pueden alinear el corpus usando modelos de alineación de palabras y entonces computar las probabilidades de traducción basdas en las palabras y los temas.

2. Introducción al LDA.

En el contexto de acceso y uso de la web, una tarea importante es revelar patrones intrínsecos del usuario que navega por la web. Este tipo de conocimiento de uso puede ser descubierto por una amplia gama de métodos estadísticos, aprendizaje de máquinas y los algoritmos de minería de datos. Entre estas técnicas, la técnica LSA basada en un enfoque de inferencia de probabilidad es uno de los paradigmas mas prometedores, que no sólo puede revelar las correlaciones subyacentes ocultas en las observaciones coocurrentes en la Web, sino que también puede identificar el factor latente de la tarea asociada con el uso del conocimiento o de la información. A continuación vamos a explorar un nuevo paradigma basado en LSA, llamado el modelo Latent Dirichlet Allocation (LDA), que se puede utilizar para el acceso y la recomendación de uso de la Web.

2.1. El Modelo Latent Dirichlet Allocation

2.1.1. Descripción general

LDA es un modelo generativo, lo que significa que trata de describir cómo se crea un documento. Se trata de un modelo probabilístico, ya que dice que un documento se crea mediante la selección de los temas y las palabras de acuerdo a las representaciones probabilísticas del texto natural. Por ejemplo, las palabras que se utilizan para escribir este párrafo se refieren a un subtema de

este documento como un todo. Las palabras reales que se usan y lo componen son elegidas en base a ese tema. La probabilidad inherente en los modelos de selección de cada palabra se deriva del hecho de que el lenguaje natural nos permite utilizar múltiples palabras dierentes para expresar la misma idea. Expresar esta idea en el modelo LDA, sirve para crear un documento sin un corpus, lo que se podria determinar cómo una distribución de temas. Para cada palabra del documento que se está generando, se escoge un tema de una distribución de Dirichlet de temas. A partir de ese tema, se coge una palabra elegida al azar basada en otra distribución de probabilidad condicionada en ese tema. Esto se repite hasta que el documento se ha generado.

La idea básica que está detrás del modelo de un corpus con una distribución Dirchilet sobre temas es que los documentos tienen varios temas y estos se superpondrán. Por ejemplo, dentro de un corpus de documentos sobre la Universidad de Princeton, habrá ponencias individuales que forman parte del Departamento de Ciencias de la Computación. Es probable que haya algunas palabras que se utilizan con más frecuencia cuando se habla del Departamento de Ciencias de la Computación que de otros departamentos en el campus, tales como: computadoras, algoritmos, gráficos, datos, modelado, y las redes. Otros departamentos, como la sociología pueden tener temas donde encontremos algunas palabras tales como: género, raza, edad, economía y Redes. El modelo LDA ve esto como un todo y elige los temas a partir de ahí. Si los documentos se compararon de forma individual, podría ser el caso de que ciertos temas no fueron recogidos, y sólo cuando todo el cuerpo es visto se empiezan a notar ciertos temas. En este ejemplo, palabras como "Redes" pueden aparecer varias veces en los documentos relativos a cualquier departamento. Esencialmente, el LDA es la creación de un modelo más realista del cuerpo, y por lo tanto, los documentos individuales. Las palabras que aparecen con menos frecuencia en los documentos únicos, pero son comunes en muchos documentos diferentes probablemente es indicativo de que existe un tema común entre los documentos. Cuando se genera un resumen, la capacidad de recoger los matices de los temas del documento permiten que la información más relevante sea incluida con menos posibilidades de repetición y dar así un resumen mejor.

La supuesta clave en el LDA es que las palabras siguen una hipótesis de "bolsa de palabras" - o, más bien que el orden no importa, que el uso de una palabra es ser parte de un tema y que comunica la misma información sin importar dónde se encuentra en el documento. Esta hipótesis dice que "Harry contrató a Sally" es lo mismo que "Sally contrató a Harry". En ambos casos, el conjunto de palabras es la misma junto con la frecuencia de cada palabra. Este supuesto es necesario para que las probabilidades sean intercambiables y que permitan una mayor aplicacion de métodos matemáticos. A pesar de que de vez en cuando trata frases semánticamente dierentes como la misma cosa, funciona bien en un documento general.

Antes de presentar el algoritmo basado en el modelo LDA para el uso y el acceso de la Web y la recomendación Web, primero miraremos un poco la evolución de los modelos generativos.

2.1.2. Modelos Generativos

Basandonos en el modelo de uso de los datos en la Web construido en forma de vector de peso durante el espacio de las páginas, entonces se crea la intención de desarrollar un mecanismo para conocer las propiedades subyacentes de los datos de uso y extraer el conocimiento de la conducta informativa de acceso web a los patrones del modelo de acceso de los usuarios. Antes de presentar el modelo LDA para el uso y acceso de la Web, es necesario recordar primero los diferentes modelos analíticos utilizados para la co-ocurrencia de las observaciones en el contexto de la minería de textos. Aunque estos modelos de análisis de datos se propusieron inicialmente para revelar la vinculación intrínseca entre los documentos y las palabras, es adecuado y razonable para introducir la idea básica en ciertos problemas de investigación, que nos ayudan a comprender fácilmente los fundamentos teóricos, así como las fortalezas de las propuestas técnicas, para llevar a cabo las tareas necesarias en el contexto de uso y acceso de la Web.

En la actualidad, existen en general dos tipos de técnicas de aprendizaje automático que pueden realizar las tareas, denominadas el modelo generativo y el modelo discriminativo. En el modelo generativo, descubrimos el modelo de la fuente a través de un procedimiento de generación, mientras que en el modelo descriptivo se aprende directamente el resultado deseado a partir de los datos de entrenamiento. En este estudio, vamos a aplicar el modelo generativo para extraer conocimiento de uso de la Web.

Lo que se pretende es introducir un modelo generativo recientemente desarrollado llamado Latent Dirichlet Allocation (LDA), y explorar la manera de emplear el modelo en la relación subyacente entre los datos de uso. El LDA es una especie de generador de modelos probabilísticos que son capaces de generar con eficacia infinitas secuencias de muestras de acuerdo a una distribución de probabilidad. El algoritmo de inferencia de probabilidad es luego usado para capturar la propiedad total de sesiones de usuario o las páginas web asociadas con los patrones de acceso de los usuarios, y revelar la semántica del tema a través de una distribución de derivados de las sesiones de usuario o los objetos de página en el espacio de trabajo latente de manera implícita.

El modelo generativo, a veces llamado el modelo Mezcla Gauss (Gaussian mixture model), puede ser usado en general para representar a las sesiones de usuario a través de un vector de expresión. En este modelo, cada visita/sesión de usuario se considera que es generada por una mezcla de temas, donde se representa cada tema por una distribución de Gauss, con una media y el valor de la variación. Los parámetros de la media y la variación se calculan mediante un algoritmo EM.

Al igual que los modelos de lenguaje de recuperación de información, donde las palabras se modelan como la co-ocurrencia de un documento, se tiene la intención de formular los accesos de usuarios o la duración dedicada a las diferentes páginas como una ocurrencia entre la sesión y la página. Aquí cada sesión de usuario consiste en una serie de páginas Web ponderadas, que se consideran equivalentes a un documento mientras que todo el conjunto de páginas web se

trata de manera similar a una "bolsa de palabras" en concepto de minería de texto.

El modelo más simple de probabilidad en la minería de textos es el modelo de unigramas, donde la probabilidad de que cada palabra es independiente de otras palabras que ya han aparecido en el documento. Este modelo es considerado como una distribución de probabilidad única U sobre todo un vocabulario V, es decir, un vector de probabilidades, U (v) para cada v palabra en el vocabulario.

Bajo el modelo de unigramas, las palabras que aparecen en todos los documentos se tomarán al azar a partir de una bolsa de palabras, y entonces sus valores son estimados. Por lo tanto, la probabilidad de una secuencia observada de la palabras w = w1, w2, ... wn es la siguiente:

$$P_{uni}(w) = \prod_{i=1}^{n} U(w_i)$$

La principal limitación del modelo de unigramas es que supone que todos los documentos son sólo colecciones de palabras homogéneas, es decir, todos los documentos presentan un solo tema, que en teoría es modelado como la distribución de probabilidad U. Sin embargo, esta suposición no es a menudo verdad en tiempo real, ya que por lo general, la mayoría de los documentos están en relación con más de un tema, que estaría representado por un conjunto marcadamente distintivo en las distribuciones. Especialmente, en el contexto del patrón de la minería el acceso de usuarios, casi todos los visitantes tienen diferentes preferencias en lugar de sólo una intención.

Con el fin de manejar la propiedad heterogénea de documentos, el modelo de mezcla se introduce para resolver el problema anterior. En este modelo generativo, en primer lugar se elige un tema, z, de acuerdo con una distribución de probabilidad, T, y luego, en base a este tema, se seleccionan las palabras de acuerdo a la distribución de probabilidad del tema. Del mismo modo, la probabilidad de observar una secuencia de palabras w = w1, w2, ... wn se formula como:

$$P_{mix}(w) = \sum_{z=1}^{k} T(z) \prod_{i=1}^{n} U_z(w_i)$$

El principal inconveniente con el modelo de mezcla es que se sigue considerando cada documento como homogéneo a pesar de que se podía hacer frente a la heterogeneidad de las colecciones de documentos. Un problema similar se produce en el contexto de la minería de uso de la Web. Por tanto, es necesario para desarrollar un mejor modelo hacer frente a la naturaleza de la heterogeneidad de las colecciones de documentos, es decir, distribuciones de probabilidad de varios temas.

El Probabilistic Latent Semantic Analysis (PLSA) es un modelo apropiado que es capaz de manejar la propiedad de varios temas en el proceso de texto Web

o el uso y acceso de la web. En este modelo, para cada palabra que observamos se recoge un tema de acuerdo a una distribución, T, que es dependiente en el documento. Los modelos de distribución de la mezcla de temas para un documento específico y cada tema, se asocia con una distribución de probabilidad sobre el espacio del vocabulario de la palabra y el cuerpo del documento, derivado de un proceso generativo. La distribución de probabilidad de una secuencia observada w = w1, w2, ... wn tiene parámetros como:

$$P_{plsa}(w) = \prod_{l=1}^{n} (\sum_{z=1}^{k} T_{w_l}(z) U_z(w_l))$$

Hay dos problemas principales con el modelo PLSA:

1. Es difícil estimar la distribución de probabilidad de un documento inédito,
2. Debido al crecimiento lineal en los parámetros que dependen del propio documento, el modelo PLSA sufre de los problemas de exceso de ajuste y de la semántica generativa inapropiada.

Para abordar estos problemas, se introduce el Latent Dirichlet Allocation (LDA) mediante la combinación de los modelos generativos básicos con una probabilidad a priori sobre los temas, ofreciendo un completo modelo generativo para los documentos. La idea básica del LDA es que los documentos son modelados como mezclas al azar sobre temas latentes con una distribución de probabilidad, donde cada tema se representa por una distribución mas un vocabulario de palabras. En este sentido, cualquier distribución mezclada al azar, T(z), está determinada por una distribución subyacente de lo que representa una incertidumbre sobre un particular $\vartheta(\cdot)$ como $p_k(\vartheta(\cdot))$, donde p_k se define sobre todo $\vartheta \in P_k$, el conjunto de todas las posibles (k-l). Es decir, los parámetros de Dirichlet determinan la incertidumbre, que contempla la distribución de la mezcla al azar sobre temas semánticos.

2.1.3. Latent Dirichlet Allocation

LDA es un modelo generativo probabilístico de observaciones co-ocurrentes. En cuanto a una aplicación típica en el análisis de textos, se utiliza el corpus de documentos como las observaciones co-ocurrentes para llevar a cabo la siguiente formulación. La idea básica del modelo LDA es que los documentos en el corpus se representan como una mezcla al azar sobre los temas latentes y cada tema se caracteriza por una distribución de las palabras en los documentos. Las notaciones utilizadas y el procedimiento generador del modelo LDA son:

Notaciones:

- M: número de Documentos.
- K: número de temas.

- V: el tamaño del vocabulario.
- α,β: parametros Dirichlet.
- ϑ_m: el tema asignado al documento m.
- $\Theta = \vartheta_m$,m = 1, ...,M: las estimaciones del corpus del tema, una matriz M x K.
- φ_k : la distribución de palabras del tema K.
- $\Phi = \varphi_k$, k = 1, ...,K: las asignaciones de la palabra de los temas, una matriz K x V.
- Dir y Poiss son funciones de distribucion Dirichlet y Poisson respectivamente.

Algoritmo de generación de un Proceso LDA:

---

for cada tema

muestra la mezcla de las palabras $\varphi_k \sim \text{Dir}(\beta)$

end

for cada uno de los documentos m = 1 : M

muestra la mezcla de los temas $\vartheta_m \sim \text{Dir}(\alpha)$

muestra la longitud de los documentos $N_m \sim \text{Poiss}(\xi)$

for cada palabra n = 1 : N_m en el documento m

muestra el índice del tema $z_{m,n} \sim \text{Mult}(\vartheta_m)$

muestra el peso de la palabra $w_{m,n} \sim \text{Mult}(\varphi_{z_{m,n}})$

end

end

El modelo generativo es, entonces, expresado de la siguiente manera:

- escoger una distribución de la mezcla ϑ $(\cdot)$ de P_k con una probabilidad P_k (ϑ)
- Para cada palabra
- Elija un tema z con una probabilidad ϑ (z).
- Elija una palabra W_i del tema z con una probabilidad T_z (W_i).

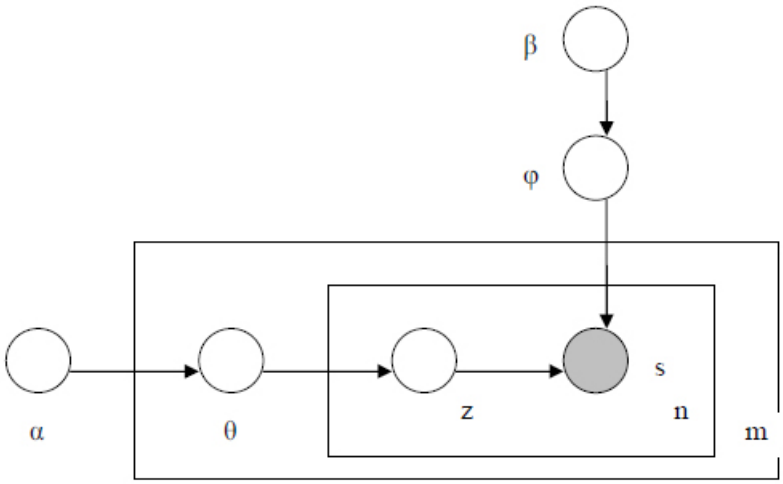
La probabilidad de observar una secuencia de palabras, w = w1, w2, ... wn, en este modelo es:

$$P_{LDA}(w) = \int_{P_k} \left\{ \prod_{i=1}^{n} \sum_{z=1}^{k} \theta(z) T_z(w_i) \right\} p_k(\theta)\, d\theta$$

donde

$$p_k(\theta) = \Gamma\left(\sum_{z=1}^{k} \alpha_z \right) \prod_{z=1}^{k} \frac{\theta(z)^{\alpha_z - 1}}{\Gamma(\alpha_z)}$$

Esta es la distribución de Dirichlet con parámetros α1, α2 ... αk. En este modelo, el objetivo final es estimar los parámetros de la distribución de Dirichlet y los parámetros para cada uno de los k modelos de los temas. A pesar de que la integral de esta expresión es intratable para una inferencia exacta, T_z (W_i) en realidad se calcula mediante el uso de una amplia gama de algoritmos de inferencia de aproximación, tales como el algoritmo de inferencia variacional. La representación gráfica del modelo LDA se ilustra a continuación:



En el LDA, un documento $d_m = \{w_{m,n}, n = 1, \ldots, N_m\}$ es generado por escoger una distribución en los temas de una distribución de Dirichlet (Dir (α)). Y teniendo en cuenta la distribución del tema, recogemos la asignación del tema para cada palabra específica. A continuación, la asignación de tema para cada marcador de posición de la palabra [m, n] se calcula por muestreo de un tema en particular de la distribución multinomial de $z_{m,n}$. Y, por último, una palabra especial de $w_{m,n}$ se genera para el marcador de posición [m, n] por muestreo del peso de la distribución multinomial de Mult ($\varphi_{z_{m,n}}$).

Como en la descripción anterior, teniendo en cuenta los parámetros α y β Dirichlet, podemos formular una distribución conjunta de un documento d_m,

una mezcla de temas de d_m, es decir, ϑ_m, y un conjunto de temas N_m, es decir, z_m es de la siguiente manera.

$$P_r(\theta_m, z_m, d_m, \Phi|\alpha,\beta) = P_r(\theta_m|\alpha) P_r(\Phi|\beta) \prod_{n=1}^{N_m} P_r\left(w_{m,n}\middle|\varphi_{z_{m,n}}\right) P_r(z_{m,n}|\theta_m)$$

A continuación, mediante la integración de ϑ_m, $\varphi_{z_{m,\,n}}$, y sumando z_m, se obtiene la probabilidad de el documento d_m:

$$P_r(d_m|\alpha,\beta) = \int\int P_r(\theta_m|\alpha) P_r(\Phi|\beta) \prod_{n=1}^{N_m} P_r\left(w_{m,n}\middle|\varphi_{z_{m,n}}\right) P_r(z_{m,n}|\theta_m) d\theta_m d\Phi$$

Por último, la probabilidad del documento del corpus $D = \{d_m, m = 1, ..., M\}$ es un producto de la probabilidad de todos los documentos del corpus.

$$P_r(D|\alpha,\beta) = \prod_{m=1}^{M} P_r(d_m|\alpha,\beta)$$

2.1.4. Estimación de parámetros Dirichlet e inferencia del tema

En general, la estimación de los parámetros del LDA se lleva a cabo mediante la maximización de la probabilidad de todos los documentos. En particular, dado un corpus de documentos $D = \{d_m, m = 1, ..., M\}$, hacemos una estimación de los parámetros α y β que maximizan la probabilidad de registro de los datos:

$$(\alpha_{est}, \beta_{est}) = \max \ell(\alpha,\beta) = \max \sum_{m=1}^{M} \log P_r(d_m|\alpha,\beta)$$

Sin embargo, el cálculo directo de los parámetros α y β es intratable debido a la naturaleza de la computación. La solución a esto es el uso de diversos métodos alternativos de estimación aproximada. Aquí empleamos el algoritmo variacional EM para estimar los parámetros de variaciones que maximizan la probabilidad total del corpus con respecto a los parámetros del modelo de α y β. El algoritmo variacional EM se describe brevemente como sigue:

- Paso 1: (paso E) Para cada documento, encontrar los valores de los parámetros de optimización variacional ϑ^*m y φ^*m.
- Paso 2: (paso M) Maximizar la banda baja del resultado de la probabilidad con respecto a los parámetros del modelo α y β. Esto corresponde a la búsqueda de la estimación de máxima verosimilitud con la posterior aproximada que se calcula en el paso E.
- El paso E y el paso M se ejecutan iterativamente hasta alcanzar un valor de máxima verosimilitud. Mientras tanto, los parámetros de estimación calculada pueden utilizarse para inferir la distribución del tema de un nuevo documento mediante la realización de la inferencia variacional.

2.2. Aplicaciones del LDA

El LDA fue desarrollado inicialmente para modelar los conjuntos de datos discretos en general, aunque sobre todo los documentos textuales. El documento original fue escrito sobre el modelo centrado en tres aplicaciones: modelos de documentos, clasificación de documentos, y filtrado de colaboración. Desde entonces, sus aplicaciones, naturalmente, han aumentado en alcance con otras investigaciones y se ha hecho mucho uso de este modelo como un marco, pero no en un contexto de resumen.

En la tarea de modelar el documento, el LDA se manejo mejor que el PLSI y una mezcla de modelos de unigramas, como era su hipótesis. El PLSI sobreajusto las probabilidades de los documentos vistos anteriormente para determinar los temas en un nuevo documento. Como era de esperar, el gran avance del LDA con el PLSI fue que fácilmente se asignan probabilidades a un documento inédito. Su aplicación en segundo lugar, con respecto a la clasificación de documentos, y observando los resultados, sugirieron que el LDA podría ser útil como un algoritmo de filtrado de velocidad para la función de selección. La última tarea que se indica fue más allá de los documentos de texto simple. El experimento EachMovie de filtrado de colaboración de datos trato de determinar las preferencias del usuario de las películas. En lugar de tener un documento de texto, tiene un usuario, y en vez de palabras sueltas, tiene películas elegidas por el usuario. El conjunto de datos se evaluó utilizando un estimador, una y otra vez y los resultados fueron que se desempeñaron mejor con el modelo LDA que con PLSI y una mezcla de unigramas.

El LDA es un modelo robusto y genérico que es fácilmente extensible más allá de los datos empíricos de un pequeño conjunto discutido. Numerosos artículos han sido publicados sobre la aplicación del LDA a una amplia gama de áreas. Se ha aplicado a las tareas que van desde la detección del fraude en las telecomunicaciones a la detección de errores en el código fuente. A pesar de la amplia gama de aplicaciones, LDA no se ha aplicado a resumen automático de documentos, aunque la posibilidad es bastante factible.

A continuación mostramos algunos de los temas en los que se ha aplicado el modelo LDA en los últimos años.

2.2.1. Utilizando LDA para descubrir patrones de acceso

Al igual que la captura de los temas subyacentes en el vocabulario de la palabra y la distribución de cada probabilidad de los documento sobre el espacio de mezcla de temas, el LDA también podría ser utilizado para descubrir temas ocultos de acceso (es decir, tareas) y las mezclas de preferencias del usuario sobre el tema del espacio cubierto a partir del historial de navegación del usuario. Es decir, a partir de los datos de uso, el LDA puede identificar los temas latentes en las páginas Web, y caracterizar a cada sesión de usuario Web como un simple de estos temas descubiertos. En otras palabras, el LDA revela dos aspectos del uso de la información subyacente, es decir, el espacio del tema oculto y la distribución de mezcla de temas de cada sesión de usuario Web, lo que refleja la

correlación subyacente entre las páginas Web, así como las sesiones del usuario Web. Con el descubrimiento de la expresión simple del tema, es posible modelar patrones de acceso de los usuarios en términos de distribución de mezcla de temas, y a su vez, para predecir las páginas potencialmente interesantes para el usuario mediante el empleo de un algoritmo de recomendación de colaboración.

El ver las sesiones de los usuarios Web en forma de mezcla de temas hace que sea posible formular el problema de la identificación de los temas/tareas subyacentes ocultas en los datos de uso. Si se dan m sesiones de los usuarios Web que expresan z temas en las páginas, podemos representar como P (p|z) a un conjunto z de distribuciones multinomiales en las n páginas, tales que P(p|z =j) = Φ(j)p , y P(z) con un conjunto de m distribuciones multinomiales sobre z temas, de tal manera que para una página en una sesion Web s, P(z =j) = Φ(s)j. Para descubrir el conjunto de temas ocultos en una colección de páginas Web p = {p_1, p_2, ... p_n} donde cada Pi aparece en algunas de las sesiones Web, nuestro objetivo es obtener una estimación de Φ que de una alta probabilidad de las páginas en la colección de las páginas. Aquí se utiliza el modelo LDA descrito anteriormente para estimar los parámetros que dan lugar a una probabilidad de registro máximo de los datos de uso. El modelo de probabilidad completa es la siguiente:

$$\theta \sim Dirichlet(\alpha)$$
$$z_i \left| \theta^{s_i} \sim Discrete(\theta^{s_i}) \right.$$
$$\phi \sim Dirichlet(\beta)$$
$$p_j \left| z_i, \phi^{z_i} \sim Discrete(\phi^{z_i}) \right.$$

En este caso, z representa un conjunto de temas ocultos, ϑ^{s_i} denota una sesión Web y la distribución de preferencias sobre los temas y Φ^{s_i} representa el tema específico de z_i de distribución de asociación sobre la colección de la página. y α y β son los hiperparámetros de ϑ y Φ. De esta manera, la ecuación se vuelve:

$$P(s_i \left| \alpha, \beta \right.) = \int p(\theta \left| \alpha \right.)(\prod_{j=1}^{n} \sum_{z_k \in Z} p(z_k \left| \theta \right.) p(p_j \left| z_k, \beta \right.))d\theta$$

Utilizamos un algoritmo de inferencia variacional para estimar la correlación de cada sesión Web con múltiples temas (α), y las asociaciones entre los temas y las páginas Web (β), con la que podemos capturar las visitas de los usuarios y su distribución de preferencias expuestas por cada sesión Web e identificar la semántica del tema. Dado un conjunto de sesiones de usuario, hacemos una estimación de los parámetros de α y β para maximizar la probabilidad de registro de los datos de uso.

El algoritmo para generar el patron de acceso a temas específicos se describe como sigue:

- [Algoritmo]: construcción de un modelo de acceso de usuario basado en el modelo LDA
- [entrada]: distribución de preferencias ϑ calculada del tema de sesión , el uso de datos SP y un umbral predefinido μ.
- [Salida]: Un conjunto de patrones de acceso de usuario AP = {apk}.
- Paso 1: Para cada tema latente z_j , seleccione todas las sesiones de usuario con $\vartheta^s_{z_j} >= \mu$ para la construcción de una agregación de sesión de usuario R_j correspondientes a Z_j
- Paso 2: Para cada tema latente z_j, calcule el patron de acceso de agregación de temas específicos de los seleccionados por el usuario en R, tomando las asociaciones de las sesiones ϑ con z_j : $ap_j = \frac{\sum_{s \in R_j} \theta^s_{z_j} \bullet s}{|R_j|}$ donde $|R_j|$ es el número de sesiones de usuario seleccionadas de R_j.
- Paso 3: salida de un conjunto de temas orientados al patron de acceso de usuarios AP sobre K temas: AP = {ap1, ap2, ... apk}.

Los paradigmas de análisis semántico latente son capaces de descubrir la relación subyacente entre las sesiones de usuario de la Web y la generación de mejores grupos de calidad en comparación con los convencionales de clustering basados en el método.

Este es un uso novedoso del enfoque de la minería Web basado en el modelo LDA. Con el modelo LDA, las asociaciones entre las sesiones de usuario y temas de navegación y las asociaciones entre los temas y la recopilación de páginas web ocultas en los registros de usuario cuando se hace click, se descubrió a través de un modelo de proceso de generación. Interpretando las páginas Web predominantes con unos resultados de probabilidades significativos en la revelación de la semántica del espacio del tema subyacente, y examinando la asociación entre sesiones de usuario y múltiples temas, se lleva a descubrir las preferencias de acceso de los usuarios sobre el espacio de tema, y a su vez, proporciona una mejor manera de identificar distintos patrones de acceso comunes mediante la agregación de las sesiones de usuario con preferencias de acceso similares. Los resultados experimentales sobre el conjunto de datos seleccionados han demostrado que la propuesta de LDA en el uso de Internet es capaz de revelar el espacio de trabajo latente y la generación de los grupos de sesión de usuario con la mejor calidad, en comparación con otros métodos convencionales basados en LSA.

2.2.2. Algoritmo de perfiles de los usuarios para la Recomendación Web sobre el modelo LDA

La idea principal de este enfoque es el uso de los pesos de las páginas dentro del espacio de trabajo dominante, sin embargo, no se necesita tener en consideración la visita histórica de los usuarios de Internet. Como consecuencia de ello, se desarrolla un algoritmo de recomendación Web a través de técnicas de filtrado basado en el modelo LDA.

Se incorpora el conocimiento descubierto en los patrones de acceso de uso con un algoritmo de filtrado colaborativo para el algoritmo de recomendación Web.

LDA es uno de los modelos generativos, que consiste en dar a conocer la correlación semántica latente entre las actividades coocurrentes a través de un procedimiento generativo. En primer lugar, se necesita conocer el patrón de uso mediante el examen de la probabilidad a posteriori de las estimaciones obtenidas a través del modelo LDA, por tanto, hay que medir las similitudes entre la sesión del usuario activo y los patrones de uso para seleccionar el perfil de usuario más encontrado, y, finalmente, hacer la recomendación de colaboración mediante la incorporación de los patrones de uso con el filtrado colaborativo, es decir, refiriéndose a las preferencias de los usuarios de otros visitantes, quienes tienen conductas similares a la navegación. Asi mismo, se cuenta con un algoritmo de ponderación que crea un sistema de puntuación en el proceso de recomendación de colaboración, para predecir las páginas de los usuarios potencialmente interesados a través de una distribución de peso de la página en el patrón de acceso más cercano.

Los resultados demuestran que el proyecto basado en la técnica LDA supera consistentemente el nivel de agrupación, la norma de agrupamiento basado en el algoritmo siempre genera una mayor precisión de recomendación y mas precisa, mientras que el rendimiento de la recomendación basado en el algoritmo PLSA se encuentra por debajo. A partir de esta comparación, se puede concluir que los enfoques basados en la recomendación propuesta en modelos de análisis semánticos latentes son capaces de hacer la recomendación web más precisa y eficaz contra los métodos convencionales de recomendación. Además de la ventaja de la alta precisión de recomendación, estos métodos también son capaces de identificar los factores semánticos latentes para que ciertos períodos de sesiones de usuario y páginas Web se agrupen en la misma categoría.

2.2.3. Un método de LDA para las preferencias selectivas

Las Preferencias selectivas codifican un conjunto de valores de los argumentos admisibles para una relación. Por ejemplo, la ubicación es probable que aparezca en el segundo argumento de la relación X se encuentra en Y, y las empresas u organizaciones en la primera. Una gran base de datos de alta calidad de las preferencias tiene el potencial de mejorar el rendimiento de una amplia gama de tareas , incluido el etiquetado de roles semánticos, la resolución de pronombres, la inferencia textual, el sentido ambiguo de una palabra, y muchos más. Por lo

tanto, se ha centrado mucha atención en la computación de forma automática a partir de un corpus de casos en relación.

Modelos sin supervisión, tales como el LDA y sus variantes se caracterizan por un conjunto de temas ocultos, que representan la estructura subyacente semántica de una colección de documentos. Estos temas ofrecen una interpretación intuitiva - que representan un conjunto de clases que almacenan las preferencias de las diferentes relaciones. Así, los modelos de temas son algo natural para el modelado de los datos.

En particular, este sistema se denomina LDA-SP y utiliza el LinkLDA (Erosheva et al., 2004), una extensión de LDA que modela a la vez dos tipos de distribuciones para cada tema. Estos dos conjuntos son los dos argumentos a favor de las relaciones. Por lo tanto, LDA-SP es capaz de capturar información acerca de los pares de los temas que comúnmente coexisten. Esta información es muy útil en la orientación de la inferencia.

Debido a que LDA-SP se basa en un modelo probabilístico formal, tiene la ventaja de que, naturalmente, se puede aplicar en muchos escenarios. Por ejemplo, podemos obtener una mejor comprensión de las relaciones similares, filtrar inferencias incorrectas sobre la base de consulta de nuestro modelo, así como producir un depósito de clase basado en las preferencias con un esfuerzo pequeño manual. En todos estos casos se obtienen resultados de alta calidad.

Se destacan dos sistemas, que aplican modelos LDA de estilo para tareas similares. OS'eaghdha (2010) propone una serie de modelos LDA de estilo para la tarea de la computación de preferencias selectivas. Este trabajo esta formado por las preferencias selectivas entre las relaciones gramaticales siguientes: verbo-objeto, nombre-nombre, y el adjetivo-sustantivo. También se centra en el modelado de la generación conjunta de ambos predicados y argumentos, y la evaluación se realiza sobre un conjunto de juicios humanos-verosimilitud para obtener resultados impresionantes en contra de Keller y Lapata (2003).

Van Durme y Gildea (2009) propuso la aplicación de LDA a las plantillas de los conocimientos generales obtenidos mediante el sistema de KNEXT (Schubert y Tong, 2003). Por el contrario, esta otra perspectiva utiliza LinkLDA y se centra en el modelado de múltiples argumentos de una relación (por ejemplo, el objeto y sujeto directo de un verbo).

Se crean una serie de modelos de tema para la tarea de la computación de preferencias selectivas. Estos modelos varían en la cantidad de independencia que se haga entre A1 y A2. En un extremo está IndependentLDA, un modelo que asume que tanto A1 y A2 se generan de forma completamente independiente. Por otro lado, JointLDA, el modelo en el otro extremo que supone que dos argumentos de una extracción específica se generan en base a una sola variable oculta z. LinkLDA se encuentra entre estos dos extremos, y es el mejor modelo para los datos de relación.

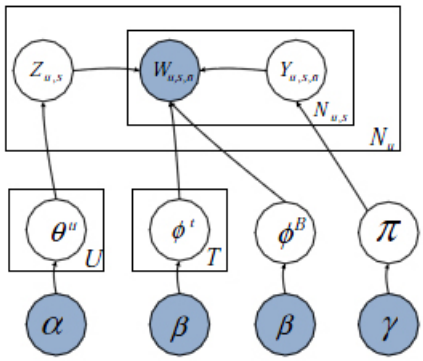
El método, LDA-SP, conforma una distribución en los temas de cada relación y al mismo tiempo agrupa las palabras relacionadas a estos temas. Este enfoque es capaz de producir clases interpretables, sin embargo, evita los inconvenientes de los enfoques basados en las clases tradicionales (mala cobertura léxica y la ambigüedad). LDA-SP alcanza el estado de la técnica de rendimiento en tareas

de predicción como la pseudo-desambiguación, y el filtrado de inferencias incorrectas. Debido a que LDA-SP genera un modelo probabilístico para completar los datos de las relaciones, sus resultados son fácilmente aplicables a muchas otras tareas como la identificación de relaciones similares, situándose en las reglas de inferencia, etc. En el futuro, se pretende aplicar este modelo para detectar automáticamente la inferencia de nuevas normas y paráfrasis.

2.2.4. Una comparación empírica con LDA de los temas en Twitter

A pesar de que también se podría aplicar LDA para descubrir temas en tweets por el tratamiento de cada tweet como un documento único, esta aplicación directa no sería muy probable que funcionara bien porque los tweets son muy cortos y a menudo contiene una sola frase. Para superar esta dificultad, algunos estudios previos proponen agregar todos los tweets de un usuario en un solo documento. De hecho, este tratamiento puede ser considerado como una aplicación del modelo tema-autor en los tweets, en el que cada documento (tweet) tiene un solo autor. Sin embargo, los temas que se descubren a veces son confusos porque los tweets agregados de un solo usuario pueden tener una amplia gama de temas. Por otro lado, este modelo no se aprovecha del siguiente punto importante: un solo tweet es por lo general sobre un tema único. Este supuesto hace uso de la restricción de longitud en Twitter. Por tanto, hay que proponer otro modelo LDA para twitter.

El modelo se basa en los siguientes supuestos. Hay T temas en Twitter, cada uno representado por una distribución de la palabra. Cada usuario tiene sus temas de interés y por lo tanto una distribución en los T temas. Cuando un usuario quiere escribir un tweet, primero elige un tema sobre la base de la distribución de tema. Luego se elige una bolsa de palabras, una por una basada en el tema elegido. Sin embargo, no todas las palabras en un tweet están estrechamente relacionados con un tema de Twitter, algunos son palabras de uso común en los tweets sobre diferentes temas. Por lo tanto, para cada palabra en un tweet, el usuario decide si es una palabra de fondo o una palabra tema y luego elige la palabra de su distribución de palabra correspondiente. Formalmente, sea φ^{t} que denota la distribución de la palabra para el tema t y φ^{B} la distribución de la palabra por las palabras de fondo. ϑ^{u} denota la distribución del tema del usuario u. π denota una distribución de Bernoulli que rige la elección entre las palabras y las palabras de fondo de tema. El proceso de generación de tweets se ilustra en la figura a continuación. Cada distribución multinomial se rige por una distribución simétrica de Dirichlet.



Se evaluó qualitativamente la efectividad del modelo Twitter-LDA en comparación con el modelo estándar LDA (es decir, el tratamiento de cada tweet en un solo documento) y el modelo tema-autor (es decir, tratar a todos los tweets de un mismo usuario en un solo documento), utilizando temas anotados manualmente. En primer lugar, se aplicó el Twitter-LDA,el LDA estándar y el modelo autor-tema para el conjunto de datos de twitter. Para el modelo estándar LDA y el modelo autor del tema, se encontran con 1000 iteraciones de muestreo de Gibbs, que es el modelo que se utiliza para calcular la inferencia. A continuación, se mezclan al azar los 330 temas de los tres modelos y se presentan las diez mejores palabras de cada tema a dos jueces humanos. Se les pidió a los jueces humanos que asignaran una puntuación a cada tema de acuerdo a las siguientes pautas: Si de la parte superior diez palabras son significativas y coherentes, se le asigna al tema una puntuación de 1, Si las diez palabras sugieren varios temas, o si hay palabras ruidosas, una puntuación de 0.5, si es imposible encontrar algún sentido fuera de las diez palabras, una puntuación de 0.

Método	Resultado	Concordancia (#encontrados/#temas)	Cohen's Kappa
Twitter-LDA	0.675	65.5%	0.433
Autor-Tema	0.539	54.5%	0.323
LDA estandar	0.509	70.9%	0.552

Podemos ver que el modelo de Twitter LDA superó claramente a los otros dos modelos. Esta comparación demuestra que el modelo Twitter-LDA es una buena opción para descubrir los temas de Twitter, todo ello basándose en el modelo LDA estándar para la creación de este nuevo modelo.

3. Análisis y síntesis de las semejanzas y diferencias entre PLSA y LDA, ventajas e inconvenientes comparativos.

3.1. Introducción

El Modelado del Lenguaje, como un enfoque estadístico para la recuperación de información, emplea la probabilidad condicional de una consulta q dado un documento d P(q|d), como una forma de clasificación por relevancia. Un enfoque particular del LM basado en IR es el PLSI. PLSI descompone la probabilidad de observar un término w y un documento d con el uso de una variable latente k así como w no se que d k. PLSI ha demostrado ser un modelo de lenguaje de baja preplejidad y supera la indexación semántica en términos de precisión y rellamada en un número de colecciones de documentos pequeño. Sin embargo, las semánticas generativas del PLSI no són completamente consistentes con lo que, hay problemas en la asignación de probabilidad para documentos previamente no observados. LDA es también un modelado de lenguaje probabilístico que posee semántica generativa consistente y supera algunas de las carencias que tiene el PLSI. Sin embargo, la siguiente sección muestra que el PLSI emerge directamente como una instancia específica del LDA así que las deficiencias del PLSI pueden ser entendidas dentro del marco del LDA.

El propósito de usar los métodos de modelado como el Latent Semantic Analisys (LSA), el Probabilistic Latent Semantic Analisys (PLSA) y el Latent Dirichlet Allocation (LDA) en el contexto de la clasificación automatizada es para reducir el ruido y para comparar las similitudes de los documentos.Se compara el desempeño de estos métodos con el fin de analizar sus diferencias en estos parámetros. Para validar la idoneidad del proceso de clasificación de ensayo, se compara con los k vecinos más cercanos (k-Nearest Neighbors, k-NN), método que se utiliza en sistemas de clasificación de ensayos.

El modelo PLSI capta la posibilidad de que un documento puede contener varios temas ya que p (z|d) es el peso de la mezcla de los temas de un documento d en particular. Sin embargo, es importante tener en cuenta que d es un índice mudo en la lista de documentos en el conjunto de entrenamiento. Por lo tanto, d es una variable aleatoria multinomial con tantos valores posibles como documentos de formación y el modelo aprende de las mezclas tema p (z |d) sólo para aquellos documentos en los que se entrena. Por esta razón, PLSI no es un modelo generativo bien definido de los documentos, no hay manera natural de que se utilicen para asignar probabilidades a un documento inédito. Otra dificultad con PLSI, que también se deriva del uso de una distribución de documentos indexados por la formación, es que el número de parámetros que deben ser estimados crece linealmente con el número de documentos de entrenamiento. Los parámetros para un modelo k-topic de PLSI son k distribuciones multinomiales de tamaño V y M mezclas sobre los k temas ocultos. Esto le da kV + kM parámetros y por lo tanto, el crecimiento lineal en M. El crecimiento lineal en los parámetros indica que el modelo es propenso a overfitting y, empíricamente,

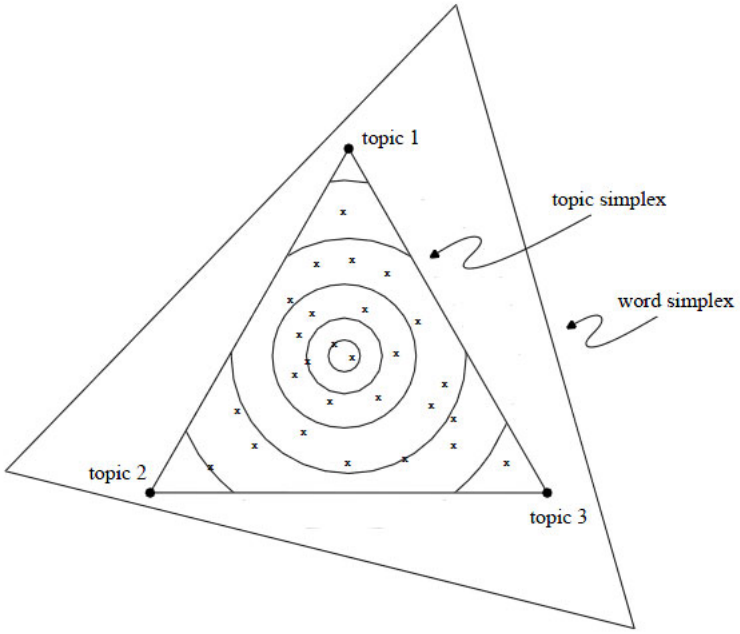
overfitting es de hecho un problema grave.

3.2. Interpretación geométrica

Una buena manera de ilustrar las diferencias entre LDA y los otro modelos latentes es considerar la geometría del espacio latente, y ver cómo un documento está representado en la geometría en cada modelo.

Los diferentes modelos como unigramas, mezcla de unigramas, PLSI, y LDA operan en el espacio de las distribuciones de las palabras. Cada distribución puede ser vista como un punto en el (V-1)-simplex, que llamamos word simplex (la palabra simple).

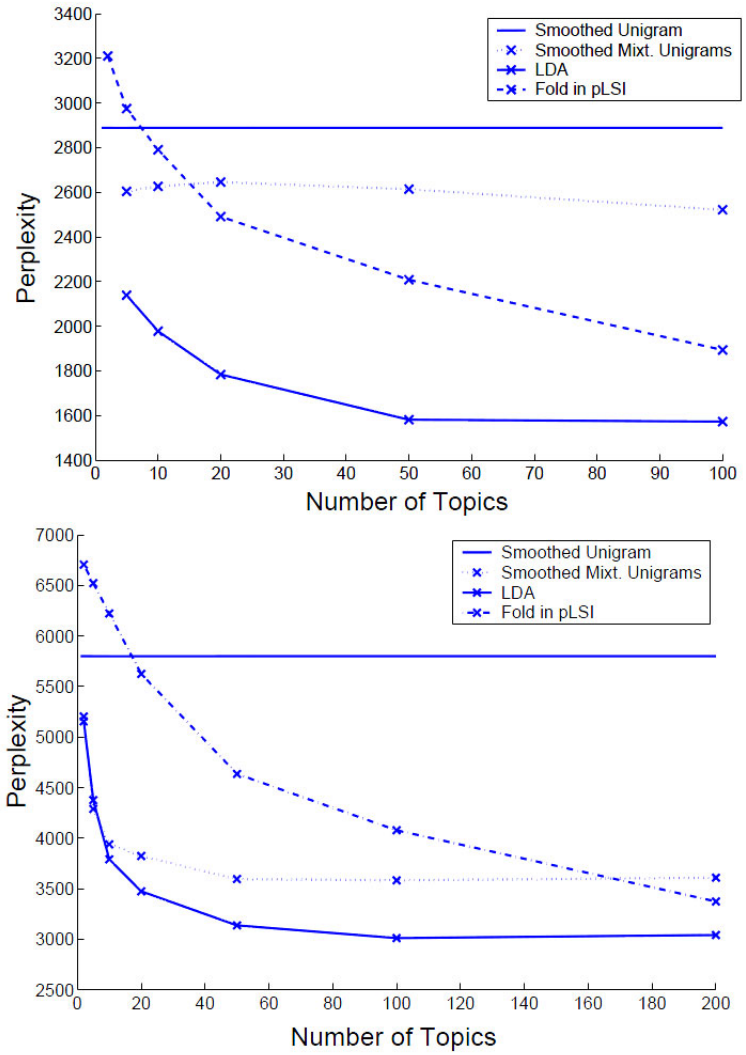
El modelo de unigramas encuentra un solo punto en el word simplex y propone que todas las palabras en el corpus provienen de su distribución correspondiente. Los modelos de variables latentes consideran k puntos en el word simplex y forman un sub-simplex basándose en esos puntos, lo que llamamos el topic simplex. Hay que tener en cuenta que cualquier punto del topic simplex también es un punto en el word simplex. Los diferentes modelos de latentes utilizan el topic simplex de diferentes maneras para generar un documento.



- El modelo PLSI propone que cada palabra de un documento de la formación proviene de un tema elegido al azar. Los temas son pintados a partir de la extracción de una distribución de documentos específicos sobre los temas, es decir, un punto en el topic simplex. Hay una distribución de este tipo por cada documento, el conjunto de documentos de capacitación por lo tanto define una distribución empírica en le simplex topic.

- El LDA, sin embargo, postula que cada palabra tanto de los documentos observados como de los ocultos es generada por un tema elegido al azar, que se extrae de una distribución con un parámetro elegido al azar. Este parámetro se muestra una vez por documento a partir de una distribución uniforme en el topic simplex.

3.3. Modelado de documento



En las gráficas anteriores se compara el LDA con unigramas, mezcla de unigramas, y los modelos de PLSI. Hemos capacitado a todos los modelos de

variables ocultas con EM con exactamente los mismos criterios de parada, que el cambio promedio en la verosimilitud esperada que es inferior a 0.001 %. Tanto el modelo de PLSI como la mezcla de unigramas sufren de problemas de overfitting graves, aunque por razones diferentes.

En el caso PLSI, el problema de agrupamiento es solventado por el hecho de que a cada documento se le permite exhibir una proporción diferente de temas. Sin embargo, PLSI solo se refiere a los documentos de capacitación y se plantea de un problema diferente de overfitting que se debe a la dimensionalidad del parámetro p (z|d). Un enfoque razonable para la asignación de probabilidades a un documento que ya no se ve es por la marginación por encima de d:

$$p(\mathbf{w}) = \sum_{d} \prod_{n=1}^{N} \sum_{z} p(w_n \mid z) p(z \mid d) p(d).$$

Este método de deducción, aunque teóricamente sólido, hace que el modelo se sobreajuste (es decir, llegué a overfit). La distribución de los temas tiene algunos componentes que están cerca de cero para aquellos temas que no aparecen en el documento. De este modo, ciertas palabras tendrán una probabilidad muy pequeña en las estimaciones de cada componente de la mezcla. Al determinar la probabilidad de un nuevo documento a través de la marginación, solo los documentos de formación que presentan una proporción similar de los temas contribuyen a la probabilidad. Para las proporciones de los temas de un documento de formación determinado, cualquier palabra que tiene una probabilidad pequeña en todos los temas constituyentes causará la perplejidad a punto de estallar. Como k se hace más grande, la probabilidad de que un documento de capacitación presente temas que abarcan todas las palabras en el nuevo documento se reduce y por lo tanto crece la perplejidad. Debemos tener en cuenta que PLSI no sobreajusta lo más rápidamente (con respecto a k) como el LDA

Este problema de overfitting está inspirado fundamentalmente en la restricción de que cada futuro documento presenta las mismas proporciones de temas que fueron vistas en uno o más de los documentos de capacitación. Dada esta limitación, no son libres de elegir las proporciones más probables de los temas para el nuevo documento. Un enfoque alternativo es el "folding-in" sugerido por Hofmann (1999), donde se hace caso omiso de la p (z|d) y reinstala los parámetros p (z|dnew). Debemos tener en cuenta que esto le da al modelo PLSI una ventaja injusta por lo que le permite volver a montar k-1 parámetros sobre los datos de la prueba.

El LDA no sufre de ninguno de estos problemas. Al igual que en PLSI, cada documento puede exhibir una proporción diferente de los temas subyacentes. Sin embargo, el LDA puede asignar probabilidades a un nuevo documento; no son necesarios sistemas heurísticos para que un nuevo documento esté dotado de un conjunto diferente de proporciones de temas que se asociaron con los documentos en el corpus de entrenamiento.

3.4. Aplicaciones del LDA y el PLSA funcionando conjuntamente.

3.4.1. Resumen de la investigación.

En el proyecto de investigación realizado por Ramesh Nallapati y William Cohen (adjuntado en las referencias de este trabajo), se trata el doble problema del descubrimiento de temas no supervisados y la estimación de influencia de temas específicos de los blogs. El modelo que proponen, intenta proporcionar al usuario blogs con los temas más interesantes para él en función de su interés particular. Comenzaron adoptando el modelo LDA al cual le añadieron una extensión que llamaron Link-LDA que definía un modelo generativo para hipervínculos y un modelo de temas específicos influenciados por los distintos documentos. Llegaron a la conclusión de que este modelo no realizaba bien su tarea de relacionar los documentos y los enlaces con una misma temática y propusieron finalmente el modelo Link-PLSA-LDA, que combina los modelos que hemos explicado y relacionado en este trabajo Juan Antonio y Jose Alberto, el PLSA y el LDA en un marco único.

El nuevo modelo, basándonos en datos de blogs, muestra una visualización de temas y blogs influentes. Además de esto, también realizaron en este estudio una evaluación cuantitativa del modelo usando logaritmos de verosimilitud de datos no vistos y en la tarea de predicción de enlaces.

Adoptan primero el modelo LDA , conocido por su eficacia en el descubrimiento de tema. Una extensión de este modelo, que se llama Link-LDA (Erosheva, Fienberg, y Lafferty 2004), define un modelo generativo de hipervínculos e influye así en tema de modelos específicos de los documentos. Sin embargo, este modelo no se aprovecha de la relación actual entre los documentos a cada lado de un hipervínculo, es decir, la idea de que los documentos tienden a vincular a otros documentos sobre el mismo tema. En este estudio proponen un nuevo modelo, llamado Link-PLSA-LDA, que combina PLSA (Hoffman, 1999) y LDA (Blei, Ng, y Jordan 2003) en un solo marco, y de forma explícita los modelos de la relación entre el tópico y la vinculación del documento enlazado. La salida del nuevo modelo en los datos revela visualizaciones de blog muy interesantes de temas y blogs influyentes sobre cada tema. También llevan a cabo una evaluación cuantitativa del modelo log-verosimilitud de los datos y no visto en la tarea de la predicción de enlace. Ambos experimentos muestran que que el nuevo modelo tiene un mejor rendimiento, lo que indica su superioridad sobre Link-LDA en temas de modelado y la influencia tema específico de los blogs.

3.4.2. Introducción

Desde hace unos años, la creación de blogs se ha extendido y se sigue extendiendo mucho en la red. Este crecimiento ha provocado que se planteen una serie de retos interesantes de investigación en la recuperación de información. Concretamente existe una necesidad de creación de técnicas automáticas para ayudar a los usuarios a acceder a blogs que realmente sean de su interés personal.

Para poder lograr este objetivo, se han estudiado los rendimientos de distintos algoritmos de clasificación como por ejemplo PageRank o HITS.

El problema que ellos abarcan, también lo estudiaron en el trabajo de Haveliwala 2002, en el cual se hizo hincapié en el PageRank de documentos precalculados para un determinado número de temas. Se basaba en la similitud de la consulta. Pero esto no era suficiente. Lo que se busca es extraer los temas de forma automática.

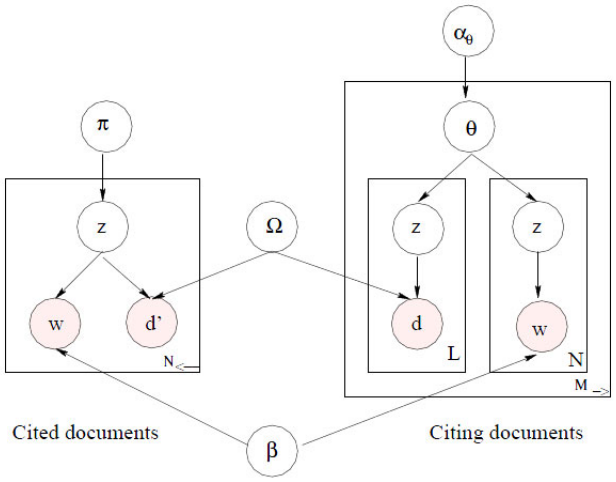
En esta investigación, el objetivo es hacer frente a ambos problemas al mismo tiempo, es decir, el descubrimiento del tema, así como la influencia de topic-models específicos de los blogs, sin supervisión. Con este objetivo, contamos con el marco de los modelos probabilísticos tema latente, como el LDA(Blei, Ng, y Jordan, 2003), y proponer un nuevo modelo en este marco. En el informe de resultados de los experimentos realizados en esta investigación sobre los datos del blog, llegan a la conclusión con algunas observaciones sobre las orientaciones para el trabajo futuro: Link-PLSA-LDA: un nuevo modelo de supervisión de los temas y la influencia de los blogs.

3.4.3. Explicación del modelo Link-PLSA-LDA

En esta sección, se describe el nuevo modelo, que se llama Link-PLSA-LDA, en detalle. La siguiente subsección presenta el proceso generativo y además describe como el modelo captura la influencia específica de los temas en los blogs. Se discuten las limitaciones del modelo.

Proceso generativo En esta investigación, los investigadores han mantenido el enfoque de Link-LDA (Erosheva, Fienberg, y Lafferty 2004) y Link-PLSA (Cohn & Hofmann 2001), en el que las citas se modelan como muestras de un tema específico de la distribución multinomial sobre los documentos citados . Por lo tanto, el proceso generador de los contenidos y las citas de los documentos que citan es el mismo que en Link-LDA. Además, con el fin de dar a conocer el flujo de información del modelo del documento donde se cita el documento citado, se ha definido un proceso explícito generativo para el contenido de los documentos citados, que hace uso de la misma distribución. En este proceso de generación nueva, los investigadores han visto el conjunto de los documentos citados como recipientes que han de cubrirse con palabras. En primer lugar, deben asociar un tema de proporciones de mezcla para el conjunto de los documentos citados. Entonces las palabras se introducen en los contenedores N $\leftarrow$ veces, donde N $\leftarrow$ es la suma de las longitudes de documento del conjunto de los documentos citados de la siguiente manera: primero se coge una muestra de un tema k de las proporciones de mezcla π, a continuación, se elige un recipiente d 'de Ω_k, y llenamos una ocurrencia de la palabra β_k en el recipiente. Este proceso es exactamente igual a la parametrización de PLSA simétrica tal como se describe en (Hoffman, 1999). Ya que utiliza una combinación de PLSA de los documentos citados y Link LDA-para citar los documentos de manera conjunta al contenido del modelo y los hipervínculos, llamando a este nuevo modelo Link-PLSA-LDA.

La representación gráfica correspondiente se muestra en la siguiente figura. Uno puede ver que la información fluye de los documentos citados en los que documentos que se citan a través de los nodos no observadosβ y Ω y, de acuerdo con el principio de separación de D en redes bayesianas (Bishop 2006).



Modelado de la influencia tema específico de los blogs Al igual que en Link-PLSA y Link-LDA, podemos interpretar $\Omega_{kd'}$ como la influencia del documento d' en el tema k. A diferencia de Link-PLSA y Link-LDA, donde esta influencia se presenta únicamente en virtud del documento d' que se cita en los documentos que tratan sobre el tema k, el nuevo modelo también tiene en cuenta el contenido de d' en la computación de la influencia actual del d'. Esto es una consecuencia directa del hecho de que también se emplea en la generación del texto de los documentos citados. Además, el parámetro k en el nuevo modelo se puede interpretar como la importancia o popularidad de cada tema en los datos. Así, el nuevo modelo nos ofrece una estadística adicional en comparación con el modelo de Link-LDA. La salida del modelo puede proporcionar al usuario que tiene blogs muy influyentes en relación con el tema de interés de los usuarios de la siguiente manera. Sea Q = (q1, ... , qn) la consulta de los usuarios que representa a sus tema de interés, uno podría devolver blogs más influyentes sobre este tema, ordenados de acuerdo a la probabilidad de los siguientes:

$$
\begin{aligned}
P(d'|Q) &= \sum_{z=1}^{K} P(d'|z)P(z|Q) \\
&\propto \sum_{z=1}^{K} P(d'|z)P(Q|z)P(z) \\
&= \sum_{z=1}^{K} \Omega_{zd'}(\prod_{i=1}^{N} \beta_{zq_i})\pi_z
\end{aligned}
$$

Mientras $\Omega_{kd'}$ representa la influencia del tema específico del documento con respecto al tema z, el termino $\prod_{i=1}^{n} \beta_{zq_i}$ representa la similitud del tema z de interés para los usuarios, mientras que π_z puede interpretarse como la importancia del tema z en el documento citado conjunto.

Las limitaciones del modelo Ya que se generan de manera diferente los documentos citados y los documentos que se citan, un solo documento no puedan tener citas y citar. Por lo tanto, el modelo supone un grafo bipartito de citas del conjunto de documentos que se citanal conjunto de documentos citados. Aunque se trata de una limitación seria de modelado, esto se puede superar fácilmente en la práctica: si un documento tiene citas y se cita también, se puede duplicar el documento, reteniendo sólo las citas de salida en un solo ejemplar y las citas de entrada en el otro y colocarlos en su respectivos conjuntos. De hecho, esta estrategia ha sido adoptada con éxito por (Dietz, Bickel, y Scheffer 2007) en su trabajo sobre las influencias de citas de modelado, que se caracteriza por una limitación similar.

Además, hay que tener en cuenta que el modelo link-PLSA-LDA define la distribución actual de las citas, en un conjunto fijo de los documentos citados. Esto significa que los nuevos documentos sólo pueden citar los documentos dentro de este conjunto fijo. Por lo tanto este modelo no es totalmente generativo, una debilidad que es compartida también por el modelo PLSA, así como el modelo de Link-LDA. Se cree que en la práctica, no es del todo descabellado suponer que el conjunto de los documentos citados se conoce en tiempo de modelado, y no va a cambiar. Por ejemplo, los documentos citados y los documentos que citan, respectivamente, podrían corresponder a los documentos ya publicados y los que se presentan actualmente en el ámbito científico, o artículos mensuales y artículos actuales de un blog.

3.4.4. Conclusiones del modelo Link-PLSA-LDA

En este trabajo se propone un nuevo modelo que descubre los temas, así como la influencia de los modelos tema específico de los blogs de una manera completamente sin supervisión. Los experimentos realizados en esta investigación demuestrán que el nuevo modelo es superior al existente Link-LDA en dos evaluaciones cuantitativas diferentes

Como parte un futuro trabajo de estos investigadores, tienen la intención de llevar a cabo experimentos en busca de palabras clave (como se describe en la sección) para evaluar el desempeño del nuevo modelos que quieren proporcionar al usuario con publicaciones en blogs muy influyentes sobre temas de su propio interés. La adquisición de la etiqueta de datos para la evaluación de blogs también forma parte de los planes de futuro.

Como se discutió en la sección, una de las deficiencias del modelo de enlace PLSA-LDA es que no es totalmente generativa. En otras palabras, el mundo de los documentos con hipervínculos es fijo y no es posible vincular a un nuevo documento en este modelo. Además, el modelo restringe el gráfico hipervínculo a una base bipartita. En la actualidad, estamos en el proceso de construcción de un nuevo modelol que es verdaderamente generativo, que permite la estructura de enlace arbitraria.

4. Conclusiones

LDA es un modelo simple, y aunque lo vemos como un competidor de métodos tales como la LSI y PLSI en el marco de reducción de dimensionalidad de las colecciones de documentos y otros aspectos discretos, también tiene la intención de ser ilustrativo de la forma en que los modelos probabilísticos pueden ser ampliados para proporcionar una maquinaria útil en los dominios de la participación de múltiples niveles de la estructura. De hecho, las principales ventajas de los modelos generativos como LDA son su modularidad y su extensibilidad.

Como un módulo probabilístico, LDA puede ser fácilmente integrado en un modelo más complejo. una propiedad que no es posible por LSI. En un trabajo reciente se han utilizado pares de módulos de LDA para modelar las relaciones entre las imágenes y sus títulos descriptivos correspondientes (Blei y Jordania, 2002). Por otra parte, existen numerosas posibilidades de extensión del LDA. Por ejemplo, LDA puede extenderse fácilmente a datos continuos u otros datos no multinomiales.

Otra simple extensión del LDA viene de permitir mezclas de distribuciones de Dirichlet en el lugar de la Dirichlet única de LDA. Esto permite una estructura más rica en el espacio de temas latentes y, en particular, permite una forma de agrupación de documentos que es diferente de la agrupación que se logra a través de temas compartidos.

Referencias

[1] Mausam Alan Ritter and Oren Etzioni. A latent dirichlet allocation method for selectional preferences. *ACL '10 Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics - Department of Computer Science and Engineering, University of Washington*.

[2] Padhraic Smyth Yee Whye Teh Arthur Asuncion, Max Welling. On smoothing and inference for topic models. *UAI '09 Proceedings of the Twenty-Fifth Conference on Uncertainty in Artificial Intelligence*, 2009.

[3] Ian R. Harris Ayanendranath Basu and Srabashi Basu. *Minimum distance estimation: The approach using density-based distances.*, volume 15. 1997.

[4] James Glass Bo-June (Paul) Hsu. Style & topic language model adaptation using hmm-lda. *EMNLP '06 Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, 2006.

[5] István Bíró. Document classification with latent dirichlet allocation. *Faculty of Informatics, Eötvös Loránd University*.

[6] Freddy Chong Tat Chua. Dimensionality reduction and clustering of text documents. *Singapore Management University*, 2009.

[7] Noah Coccaro and Daniel Jurafsky. *Proceedings of ICSLP-98*, volume 6. 1998. Towards better integration of semantic predictors in statistical language modeling.

[8] D. A. Cohn and T. Hofmann. The missing link - a probabilistic model of document content and hypertext connectivity. 2000.

[9] Bryan Thompson David Guo, Michael Berry and Sidney Balin. Knowledge-enhanced latent semantic indexing. *Information Retrieval*, 2003.

[10] Andrew Y. Ng David M. Blei and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research 3*, 2003.

[11] Chris H. Q. Ding. A similarity-based probability model for latent semantic indexing. *Proceedings of SIGIR-99*, 1999.

[12] Adam Berger Doug Beeferman and John Lafferty. *Statistical models for text segmentation*. 1997. Machine Learning.

[13] Damiano Spina Enrique Amigó and Bernardino Beotas. Evaluación de sistemas para la monitorización de contenidos generados por usuarios. *Departamento de Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia*.

[14] J. Ma G. Xu, Y. Zhang and X. Zhou. Discovering user access pattern based on probabilistic latent factor model. *ADC*, 2005.

[15] Lin Li Guandong Xu, Yanchun Zhang. *Web Mining and Social Networking, Techniques and Applications*. Springer, 2011.

[16] Qiang Yang Yong Yu Gui-Rong Xue, Wenyuan Dai. Topic-bridged plsa for cross-domain text classification. 2010.

[17] Marti A. Hearst. Texttiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics*, 1997.

[18] T. Hofmann. Probabilistic latent semantic analysis. *Proc. of Uncertainty in Artificial Intelligence, UAI99*, 1999.

[19] Thomas Hofmann. *Unsupervised learning by probabilistic latent semantic analysis*. 2001.

[20] Edgar Tello-Leal Victor Sosa-Sosa Isidra Ocampo-Guzmán, Ivan Lopez-Arevalo. Hacia la construcción automática de ontologías. *Laboratorio de Tecnologías de Información, Universidad Autónoma de Tamaulipas*.

[21] Jácint Szabó István Bíró and András A. Benczúr. Latent dirichlet allocation in web spam filtering. *AIRWeb '08 Proceedings of the 4th international workshop on Adversarial information retrieval on the web*, 2008.

[22] Patrick Jähnichen. *Finding and Analyzing Social Networks in unstructured web log data using probabilistic topic modeling*. PhD thesis, Natural Language Processing Group, University of Leipzig, Germany.

[23] Wen-tau Yih Jianfeng Gao, Kristina Toutanova. Clickthrough-based latent semantic models for web search. 2011.

[24] Guoliang Li Lizhu Zhou Ju Fan, Hao wu. Suggesting topic-based query terms as you type. *Web Conference (APWEB), 2010 12th International Asia-Pacific - Department of Computer Science and Technology, Tsinghua University*.

[25] Suin Kim. Latent dirichlet allocation. 2011.

[26] Michalis Vazirgiannis Magdalini Eirinaki. Web mining for web personalization. *ACM Transactions on Internet Technology (TOIT)*, Volume 3 Issue 1, 2003.

[27] David M. Blei Matthew D. Hoffman and Francis Bach. Online learning for latent dirichlet allocation. *Advances in Neural Information Processing Systems 22*, 2009.

[28] Florent Monay and Daniel Gatica-Perez. On image auto-annotation with latent space models. *MULTIMEDIA '03 Proceedings of the eleventh ACM international conference on Multimedia*.

[29] Kenton W. Murray. *Summarization by Latent Dirichlet Allocation: Superior Sentence Extraction through Topic Modeling.* PhD thesis, Department of Computer Science, Princeton University, 2009.

[30] Jian-Yun Nie. Information retrieval - lsi, plsi and lda.

[31] P. Frasconi P. Baldi and P. Smyth. *Modeling the internet and the web: probabilistic methods and algorithms.* 2003.

[32] Hang Li Nick Craswell Quan Wang, Jun Xu. Regularized latent semantic indexing. 2011.

[33] C. A. Hurtado R. A. Baeza-Yates and M. Mendoza. Improving search engines by query clustering. *JASIST*, 2007.

[34] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. *NIPS*, 2001.

[35] Martin Ester Samaneh Moghaddam. Ilda: Interdependent lda model for learning latent aspects and their ratings from online product reviews. 2011.

[36] Thomas K. Landauer GeorgeW. Furnas Scott C. Deerwester, Susan T. Dumais and Richard A. Harshman. *Journal of the American Society of Information Science.* Indexing by latent semantic analysis.

[37] Thomas K. Landauer GeorgeW. Furnas Scott C. Deerwester, Susan T. Dumais and Richard A. Harshman. Indexing by latent semantic analysis. *Journal of the American Society of Information Science*, 1990.

[38] Tom Landauer Thomas Huffman and Hsuan-Sheng Chiu Haiyan Qiao Jonathan Huang Peter Foltz, Melanie Martin. Lsa, plsa, and lda.

[39] Francine Chen Thorsten Brants and Ioannis Tsochantaridis. *Proceedings of Conference on Information and Knowledge Management.* 2002. Topic-based document segmentation with probabilistic latent semantic analysis.

[40] Senya Kiyasu Tomonari Masada and Sueharu Miyahara. Comparing lda with plsi as a dimensionality reduction method in document clustering. *Large-Scale Knowledge Resources, Construction and Application, Third International Conference, LKR 2008.*

[41] Erkki Sutinen Tuomo Kakkonen, Niko Myller and Jari Timonen. Comparison of dimension reduction methods for automated essay grading. *Technology and Society 2008 - www.mendeley.com - Department of Computer Science and Statistics, University of Joensuu, Finland*, 2008.

[42] Cory Reina Usama M. Fayyad and Paul S. Bradley. Initialization of iterative refi nement clustering algorithms. *Knowledge Discovery and Data Mining*, 1998.

[43] Xing Wei and W. Bruce Croft. Lda-based document models for ad-hoc retrieval. *SIGIR '06 Proceedings of the 29th annual international ACM SIGIR Conference - Computer Science Department, University of Massachusetts Amherst.*

[44] Huiwen Wu and Dimitrios Gunopulos. Evaluating the utility of statistical phrases and latent semantic indexing for text classifi cation. *Proceedings of IEEE International Conference on Data Mining*, 2002.

[45] Y. Zhou X. Jin and B. Mobasher. A unified approach to personalization based on probabilistic latent semantic models of web usage and content. *Proceedings of the AAAI 2004 Workshop on Semantic Web Personalization (SWP 04)*, 2004.

[46] Qiming Luo Hui Xiong Xiaojun Quan, Enhong Chen. Adaptive label-driven scaling for latent semantic indexing. 2010.

[47] Jing Jiang Xin Zhao. An empirical comparison of topics in twitter and traditional media. *Singapore Management University, School of Information Systems*, 2011.

[48] Guandong Xu. Web mining techniques for recommendation and personalization. *Faculty of Health, Engineering & Science, Victoria University, Australia*, 2008.

[49] Edward Y. Chang Zhiyuan Liu, Yuzhou Zhang. Plda+: Parallel latent dirichlet allocation with data placement and pipeline processing. *ACM Transactions on Intelligent Systems and Technology (TIST)*, volume 2 Issue 3, 2011.