Improving Social Bookmark Search Using Personalised Latent Variable Language Models

Morgan Harvey and Ian Ruthven[1], Mark J. Carman[2]

[1]University of Strathclyde [2]University of Lugano

Lukas Zorich (27 de octubre, 2016)

Introducción

- Los sistemas de etiquetado social (*Social Tagging Systems*) permiten al usuario anotar cada recurso con cualquier tipo de *tag*.
- El resultado es una categorización personalizada definida por los propios usuarios.

Problema

- El problema es que las etiquetas varían mucho entre usuarios, resultando en un gran número de *tags* polisémicos y sinónimos.
- Esto hace que la búsqueda de recursos en estos sistemas sea muy difícil.

Solución

- Una forma de mejorar la búsqueda es agrupar de alguna forma los términos que tengan un significado parecido.
- Otra forma, es personalizar la búsqueda basado en las preferencias del usuario y sus intereses. Esto se puede hacer implícitamente utilizando los recursos que el usuario guardó o marcó y las etiquetas que ocupó en estos marcadores.

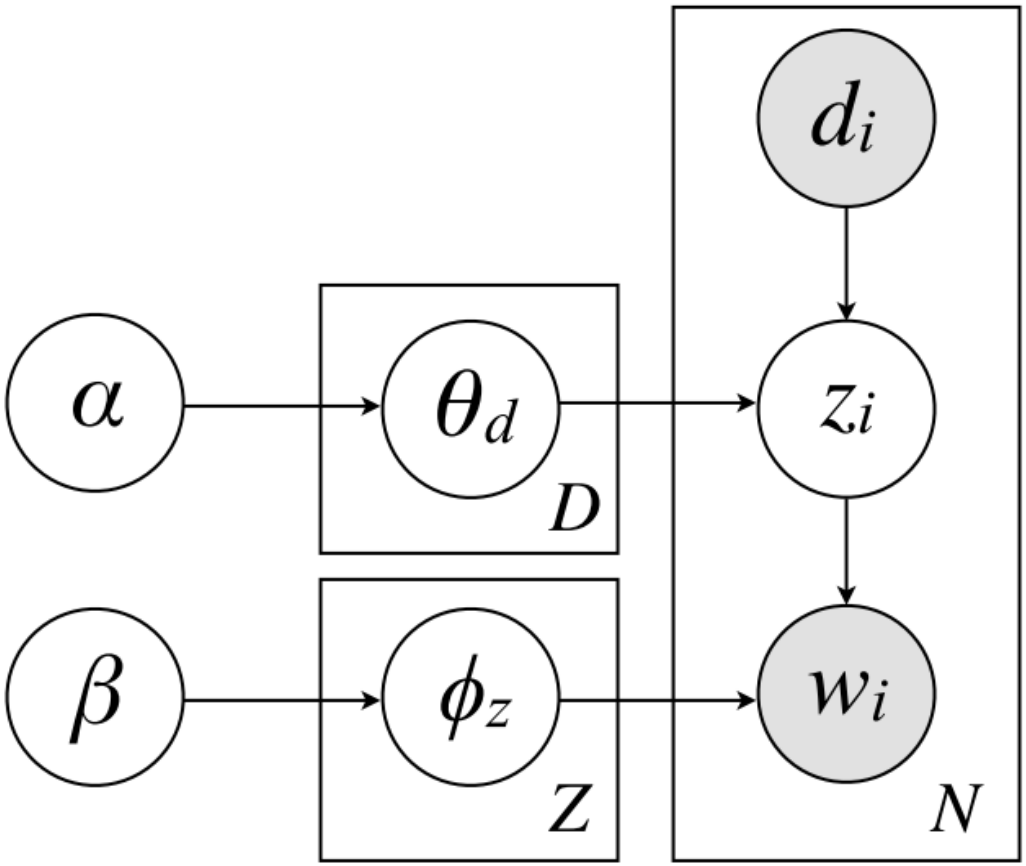
Topic Models

- El modelado de tópicos permite encontrar la estructura semántica de una colección de documentos en forma de tópicos.
- En este caso en particular, los "documentos" serán las URLs que el usuario marcó.
- La representación del documento está dada por todos los *tags* de todos los usuarios que etiquetaron ese marcador.
- Esto permitirá:
  1. Manejar los sinónimos y la polisemia.
  2. Saber cuando dos recursos (o URLs) son similares.

Latent Dirichlet Allocation (LDA)

- LDA es el modelo más conocido a la hora de modelar tópicos.
- Modelo probabilístico que representa cada documento como una mezcla de categorías o tópicos, y a su vez, cada tópico es una mezcla de palabras del vocabulario.

Latent Dirichlet Allocation (LDA)



Latent Dirichlet Allocation (LDA)

- Los parámetros que hay que estimar son θ_d y ϕ_z, que contienen la probabilidad de una palabra dado un tópico $P(w \mid z)$ y un tópico dado un documento $P(z \mid d)$.
- Con los parámetros estimados, podemos calcular por ejemplo qué documentos o palabras son similares entre sí.

Tagging Topic Model 1 (TTM1)

- En LDA, se construyen distribuciones de los tópicos sobre los recursos (URLs) $P(z \mid d)$, que se puede entender como los tópicos más probables de cada recurso.
- En *social tagging systems* se debería considerar también distribuciones sobre los usuarios que indiquen los tópicos que el usuario prefiera.
- Asumiendo que la probabilidad de un usuario y un recurso son independientes dado un tópico, se define $\theta_d = P(z \mid d)$ (como en LDA) y $\psi_u = P(z \mid u)$.

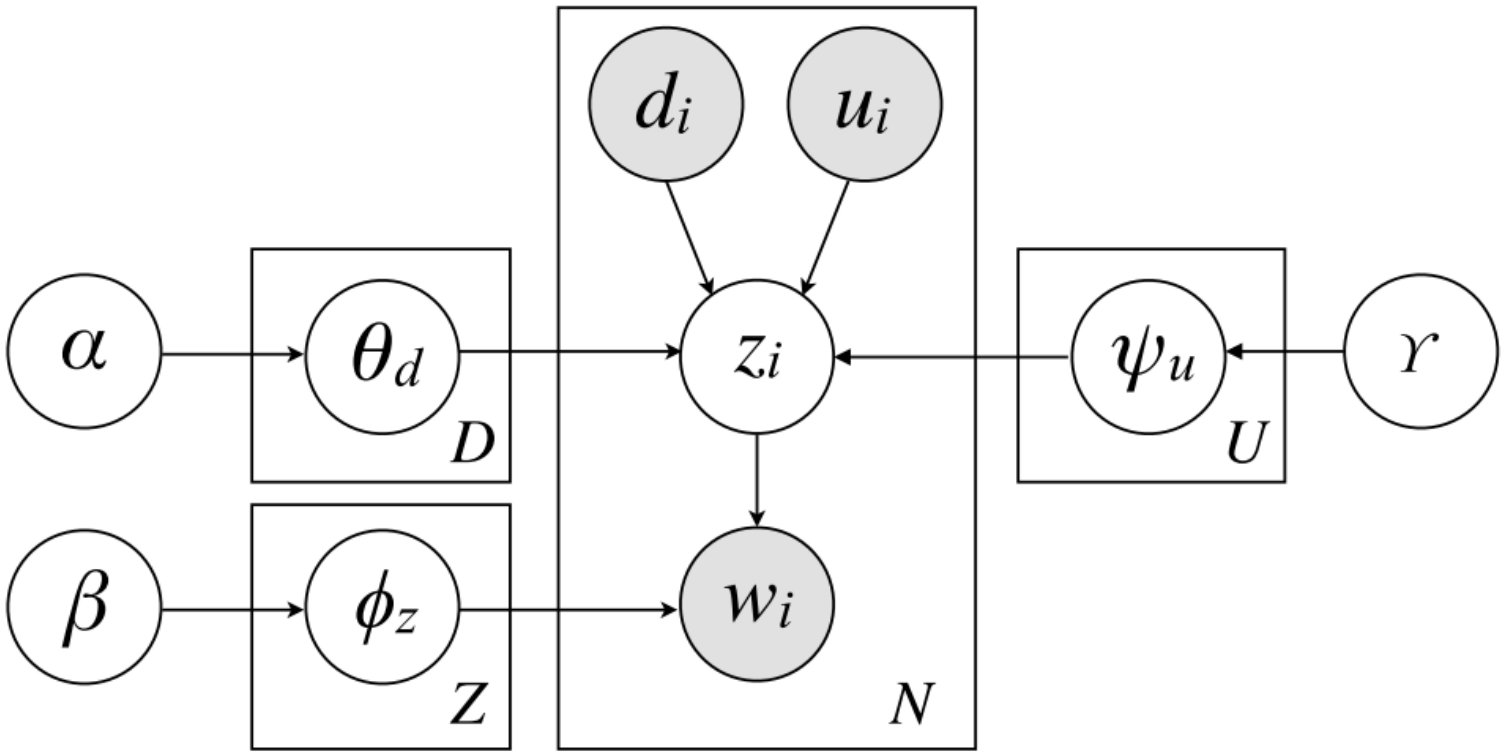
Tagging Topic Model 1 (TTM1)

- La probabilidad de un tópico dado un usuario u y un recurso d es:

$$P(z \mid \theta_d, \psi_u) \propto \frac{P(z \mid \theta_d) P(z \mid \psi_u)}{P(z)}$$

- Se pone un *prior* Dirichlet de parámetro γ sobre la distribución de tópicos por usuario ψ_u

Tagging Topic Model 1 (TTM1)



Tagging Topic Model 2 (TTM2)

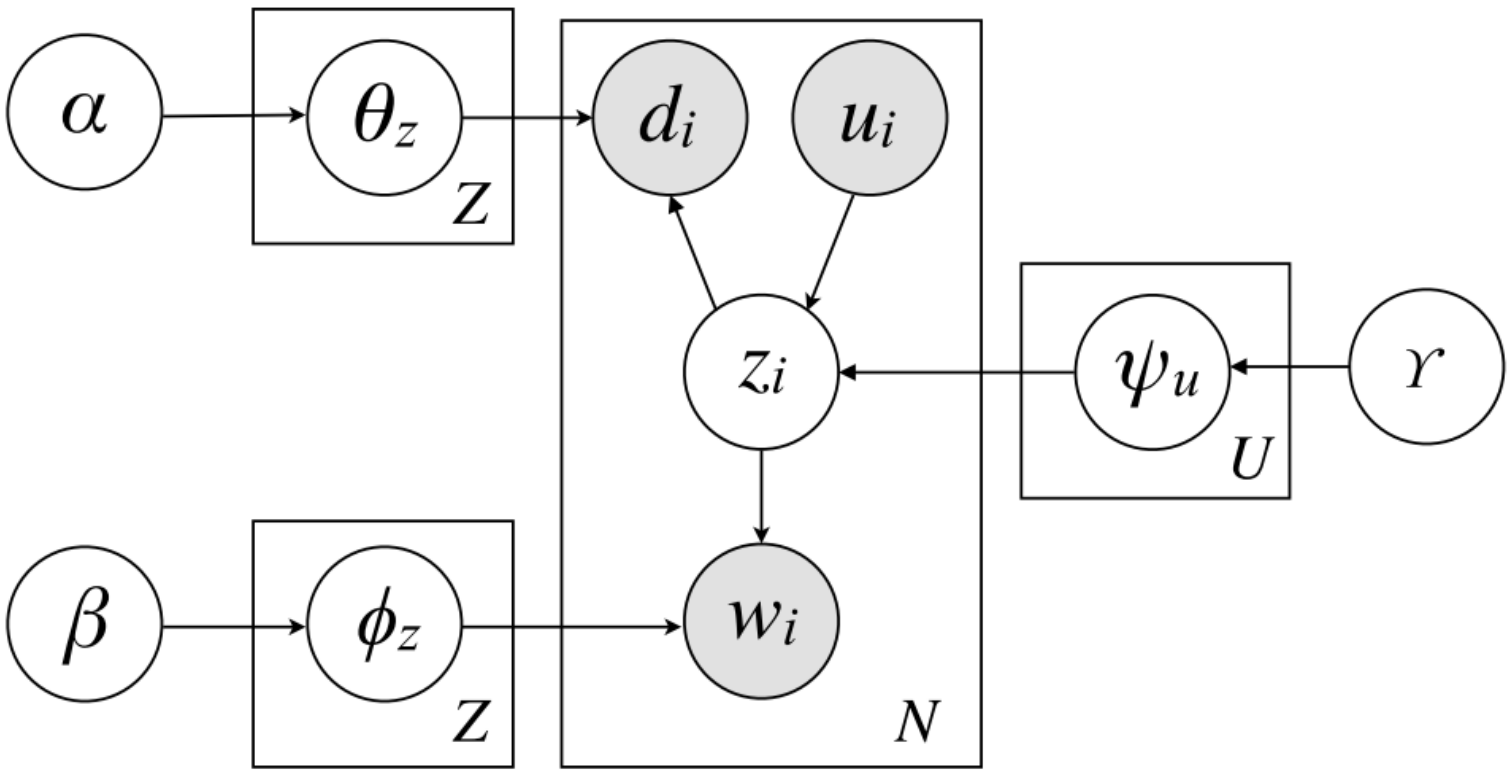
- La generación de *tags* sociales no calza con el modelo TTM1, ya que en este modelo se asume que cada documento "elige" la distribución de tópicos θ_d
- En la vida real, el usuario elige un tópico en el que está interesado y luego en base a este tópico busca recursos para marcar y etiquetar.
- Se propone el modelo Tagging Topic Model 2 (TTM2), donde el recurso es elegido por el tópico y no al revés.

Tagging Topic Model 2 (TTM2)

Para cada $i = 1 \ldots N$,

1. Se elige aleatoriamente un tópico z_i de la distribución de tópicos del usuario u, $P(z \mid u)$.
2. Se elige aleatoriamente un recurso d_i de la distribución de documentos del tópico z_i, $P(d \mid z)$
3. Se elige aleatoriamente el *tag* w_i de la distribución de *tags* $P(w \mid z)$ del tópico z_i.

Tagging Topic Model 2 (TTM2)



Ranking Resources

- Dado una query q se quiere retornar al usuario un conjunto de recursos ($d \in D$) rankeados.
- En el caso de LDA, la fórmula del ranking sería:

$$P(d \mid q) \propto P(d)P(q \mid d) = P(d) \prod_{w \in q} P(w \mid d)$$
$$= P(d) \prod_{w \in q} \sum_{z} P(w \mid z)P(z \mid d)$$

Ranking Resources

- En el caso de los modelos TTM, también se sabe el usuario que ingresó la *query*, por lo que sus preferencias también pueden ser ingresadas en el ranking.
- Para el modelo TTM1, se calcula el ranking como:

$$P(d \mid q, u) \propto P(d \mid u)P(q \mid d, u) = P(d \mid u) \prod_{w \in q} P(w \mid d, u)$$

donde,

$$P(d \mid u) = P(d) \sum_{z} \frac{P(z \mid d)P(z \mid u)^{\pi_u}}{P(z)}$$

$$P(w \mid d, u) = \frac{\sum_z P(w \mid z)P(z \mid d)P(z \mid u)^{\pi_u} P(z)^{-1}}{\sum_z P(z \mid d)P(z \mid u)^{\pi_u} P(z)^{-1}}$$

Ranking Resources

- En el modelo TTM2, $P(d \mid u)$ y $P(q \mid d, u)$ son:

$$P(d \mid u) = \sum_z P(d \mid z) P(z \mid u)^{\pi_u}$$

$$P(w \mid d, u) = \frac{\sum_z P(w \mid z) P(d \mid z) P(z \mid u)^{\pi_u}}{\sum_z P(d \mid z) P(z \mid u)^{\pi_u} P(z)}$$

Experimentos: Dataset

Datos sacados de Delicious:

Metric	Original	Reduced
users	60,663	9,587
URLs	476,248	111,232
vocab count	113,428	14,023
bookmarks	3,235,299	569,117
word occurrences	12,294,136	2,473,738
avg bookmarks/user	53.3	59.4
avg bookmarks/URL	6.79	5.1
avg annotations/URL	25.8	22.2
avg annotations/bookmark	3.8	4.3

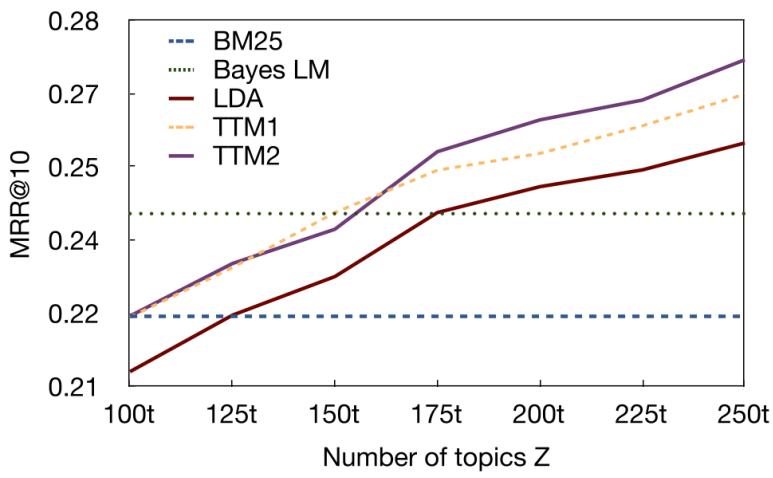
Experimentos: Evaluación

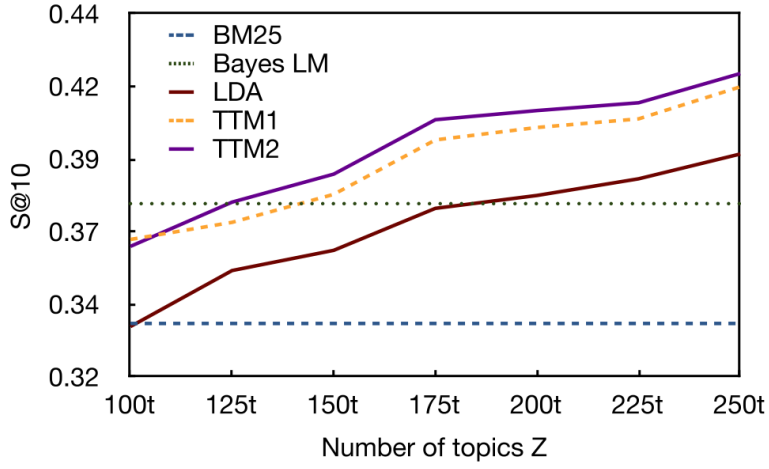
- Se seleccionaron los últimos 10% de *bookmarks* de cada usuario para el *testing*.
- Se usaron los *tags* de los *bookmarks* de *testing* como *query*.
- Se clasifica un recurso rankeado como relevante si el usuario que hace la *query* efectivamente marcó ese documento.
- Métricas:
  $\text{S@k} = \frac{1}{|q|} \sum_i^{|q|} I(rank(d_i, q_i) <= k)$
  $\text{MRR} = \frac{1}{|q|} \sum_i^{|q|} \frac{1}{rank(d_i, q_i)}$
- *Baselines*: SMatch, BM25 y BayesLM.

Resultados

	S@1	S@5	S@10	MRR@10
SMatch	0.0555	0.1372	0.1860	0.0900
BM25	0.1701	0.2975	0.3376	0.2238
BayesLM	0.1819	0.3299	0.3772	0.2440
LDA	0.1994	0.3397	0.3936	0.2579
TTM1	0.2030	0.3556∗	0.4158∗	0.2675∗
TTM2	**0.2137**†	**0.3559**∗	**0.4202**∗	**0.2743**†

Resultados: Sensibilidad al número de tópicos





Resultados: Sensibilidad al número de bookmarks

	S@10		**MRR@10**	
	0-60	**60-80**	**0-60**	**60-80**
SMatch	0.1707	0.1667	0.0815	0.0811
BM25	0.3232	0.3271	0.2098	0.2180
BayesLM	0.3624	0.3776	0.2344	0.2291
LDA	0.3694	0.3941	0.2212	0.2534*
TTM1	0.3705	0.4175*	0.2361	0.2700*
TTM2	0.3719	0.4454*	0.2394	0.2804*