

Immersive Recommendation

News and Event Recommendations Using Personal Digital
Traces

Romano Fenzo

Contexto

Crecimiento de la web, redes sociales, e-commerce, y comunicaciones con dispositivos móviles generan continuamente *digital traces*

Problemas

Poco o nulo análisis sobre estos datos para mejorar la experiencia del usuario

Datos con **ruido** relacionado al contexto

Ej. Post de Twitter contienen muchos términos sociales como *love*, *share*, *retweet*, etc.

Múltiples fuentes de datos.

Términos

Se busca generar recomendaciones de noticias y eventos cercanos utilizando *digital traces*.

Contexto: Una fuente de datos que forma parte de las *digital traces*.

Ej. Contexto1: **Facebook**, Contexto2: **Twitter**, Contexto3: **email**.

Items: Objetos a recomendar, en este caso **noticias y eventos cercanos**.

Ej. Evento X de **Meetup.com**, Noticia X de **Medium.com**

Propuesta

Fase 1. Perfilar al usuario según tópicos

Channel-Aware Latent Dirichlet Allocation (CA-LDA)

Fase 2: Recomendación de *items*

Hybrid recommendation algorithm (Perfiles + Ratings)

Fase 1. Perfilar al usuario según tópicos

Opción común: LDA simple

Procedimiento:

1. Entrenar LDA en base a un solo contexto
2. Inferir los tópicos de un usuario para cada contexto por separado
3. Perfilar el usuario como la suma ponderada de la inferencia anterior

Fase 1. Perfilar al usuario según tópicos

Opción común: LDA simple

Problemas:

1. **Cobertura Insuficiente:** Si el texto de los *items* es pequeño, el perfilamiento será poco acertado
2. **Ruido de contexto**

Fase 1. Perfilar al usuario según tópicos

Propuesta: Channel-Aware LDA (CA-LDA)

Entrena LDA sobre la totalidad de contextos.

Lo anterior asume que:

1. Hay tópicos destacados que son compartidos por los contextos
2. Hay tópicos específicos por contexto (asociado al ruido)

Fase 1. Perfilar al usuario según tópicos

Propuesta: Channel-Aware LDA (CA-LDA)

Entrena LDA sobre la totalidad de contextos.

Lo anterior asume que:

1. Hay tópicos destacados que son compartidos por los contextos
2. Hay tópicos específicos por contexto (asociado al ruido)

Fase 1. Perfilar al usuario según tópicos

Propuesta: Channel-Aware LDA (CA-LDA)

Modificaciones a LDA:

Para cada contexto, genera una distribución aleatoria de palabras en los tópicos específicos.

Genera una única distribución aleatoria de palabras en los tópicos comunes.

Fase 1. Perfilar al usuario según tópicos

Propuesta: Channel-Aware LDA (CA-LDA)

Modificaciones a LDA:

Mediante una binomial decide, dentro del algoritmo de LDA, si la palabra a procesar es común o específica.

Genera las palabras de acuerdo al resultado anterior

Fase 1. Perfilar al usuario según tópicos

Propuesta: **Channel-Aware LDA (CA-LDA)**

Resultados:

Términos que prevalecen en determinado contexto son asignados a los tópicos específicos


Previene que los tópicos específicos se mezclen con los comunes.

Contextos con vocabulario pobre, son enriquecidos (entrenamiento en conjunto)

Fase 1. Perfilar al usuario según tópicos

Resultados:

Términos
específicos



Context	Background Terms
Email	<i>pleas, offic, schedul, convers, fax, cellular</i>
Twitter	<i>awesom, share, tweet, post, video, love</i>
Facebook	<i>love, night, happi, tomorrow, final, tonight</i>
Medium.com	<i>happen, idea, actual, experi, hard, reason</i>
Meetup.com	<i>social, event, network, singl, profession, join</i>

Nota: Adicionalmente se generan **perfiles a cada *item***

Fase 2. Recomendación de *items*

Propuesta: *Hybrid recommendation algorithm*

Utiliza perfiles creados en Fase 1 y *ratings* sobre los items

$$r_{ij} \sim \mathcal{N}((\mathbf{u}_i + \boldsymbol{\eta}_i)^T (\mathbf{v}_j + \boldsymbol{\varepsilon}_j), c_{ij}^{-1})$$

$$\boldsymbol{\eta}_i \sim \mathcal{N}(0, \lambda_u^{-1} I_K)$$

$$\boldsymbol{\varepsilon}_j \sim \mathcal{N}(0, \lambda_v^{-1} I_K)$$

r_{ij} : rating de usuario i al *item* j

\mathbf{u}_i , \mathbf{v}_j : perfil de usuario i e *item* j respectivamente

$\boldsymbol{\eta}_i$, $\boldsymbol{\varepsilon}_j$: compensan la información que no fué capturada en el perfilamiento y es otorgada por el *rating*

$c_{i,j}$: confianza sobre r_{ij} , ej. *Medium upvote* vs 1-5 *ratings*

Nota: $\boldsymbol{\eta}_i$ y $\boldsymbol{\varepsilon}_j$ son calculados mediante clustering (EM - Expectation Maximization)

Dataset

Medium.com: 63,053 users, 31,000 noticias

En promedio, cada usuario ha calificado positivo 13.1 noticias

10% de las noticias tienen el 59% de las calificaciones positivas

Meetup.com: 50,000 users, 11,823 grupos de juntas

En promedio, cada usuario pertenece a 5.1 grupos

10% de los grupos tienen el 61% de los usuarios

Dataset

Para cada usuario:

Sus 3000 tweets más recientes

Tweets de personas que ellos siguen

Tweets que ocupan *hashtags* que el usuario también usa

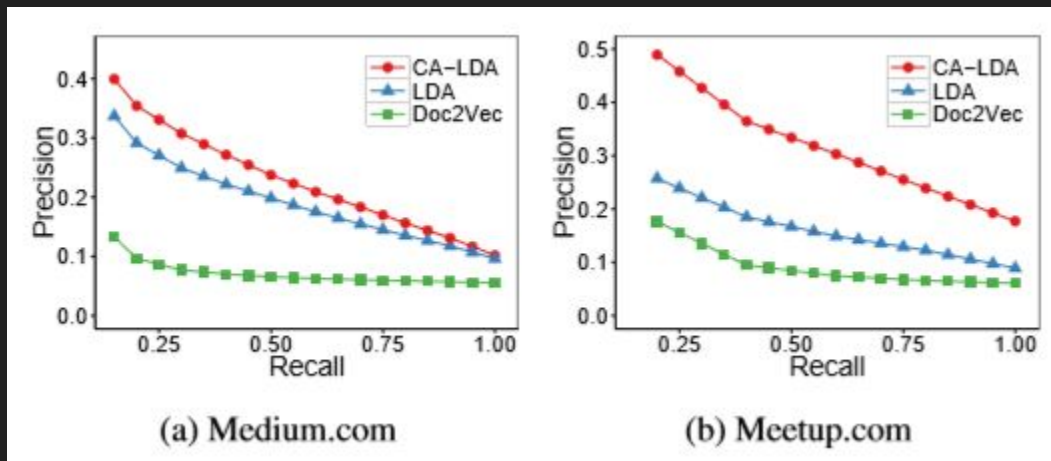
Evaluación

Grupo de test:

3000 usuarios aleatorios de Medium.com y Meetup.com

Para cada usuario se eligieron 10 noticias que calificó positivo, y 190 que no calificó, 5 grupos a los que se unió y 95 a los que no se unió.

Evaluación



Comportamiento de mean Average Precision (mAP) cuando se consideran distintas fuentes de tweets.

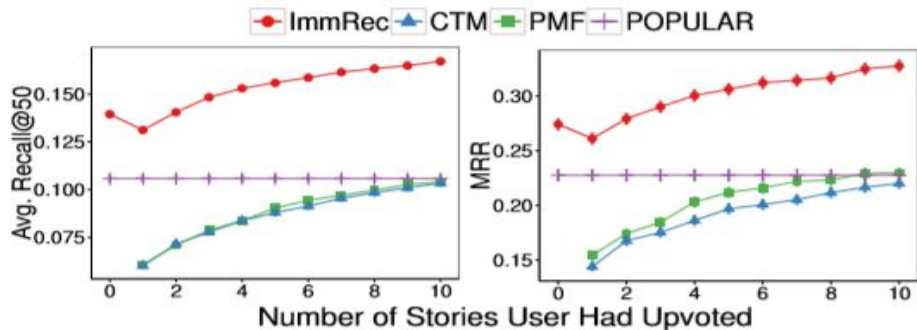
mAP	Own	Followees	Hashtags	Combined
Medium	0.237	0.240	0.188	0.245
Meetup	0.331	0.347	0.282	0.353

Evaluación

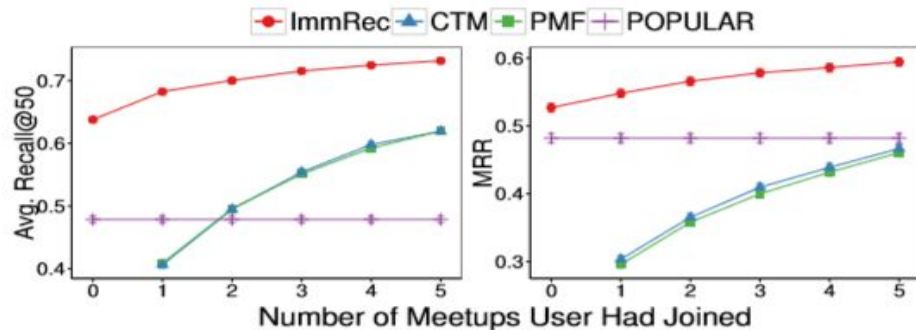
Comparativa de Immersive Recommendation con Most Popular First, Probabilistic Matrix Factorization, Collaborative Topic Modeling

Recall Promedio: (Items positivos correctos en top-M) / (Total items positivos)

MRR: Mide el ranking del primer item correcto y lo promedia sobre todos los usuarios.



(a) Medium.com



(b) Mectup.com