

# Filtrado Basado en Contenido II

IIC 3633 - Sistemas Recomendadores

Denis Parra, Ivania Donoso  
Profesor Asistente, DCC, PUC Chile

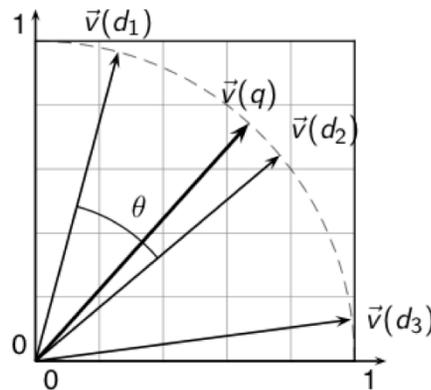
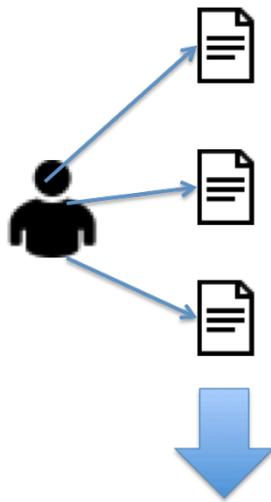
# TOC

## En esta clase

1. Representación y Aprendizaje del Modelo de Usuario
  1. Rocchio's Algorithm
  2. Revisión de Syskill & Webert
2. Tag-based recommendations
3. Actividad Práctica con R, RStudio y Packages tm() y ggplot2()

# Representación del Modelo de Usuario I

Bajo un escenario estático, podemos agregar el modelo del usuario en función de los contenidos por los cuales ha mostrado preferencias.



Doc\_1 = {w\_1, w\_2, ..., w\_3}



Doc\_2 = {w\_1, w\_2, ..., w\_3}



Doc\_3 = {w\_1, w\_2, ..., w\_3}



Doc\_n = {w\_1, w\_2, ..., w\_3}

user\_profile = {w\_1, w\_2, ..., w\_3} usando  
TF-IDF

- ¿Cómo incorporar información para actualizar el modelo?
- ¿Debiéramos "decaer" la importancia de ítems antiguos?
- ¿Distinta metadata debería integrarse en una sola bolsa de palabras o separarse?

# Representación del Modelo de Usuario II

## Modelo de Relevancia Rocchio

- Actualización de Query en base a feedback Positivo y Negativo

$$Q_{i+1} = \alpha Q_i + \beta \sum_{rel} \frac{D_i}{|D_i|} - \gamma \sum_{nonrel} \frac{D_i}{|D_i|}$$

- Clasificador que combina vectores de documentos para cada clase

$$\omega_{ki} = \beta \cdot \sum_{\{d_j \in POS_i\}} \frac{\omega_{kj}}{|POS_i|} - \gamma \cdot \sum_{\{d_j \in NEG_i\}} \frac{\omega_{kj}}{|NEG_i|}$$

# Representación del Modelo de Usuario II

## Modelo de Relevancia Rocchio II

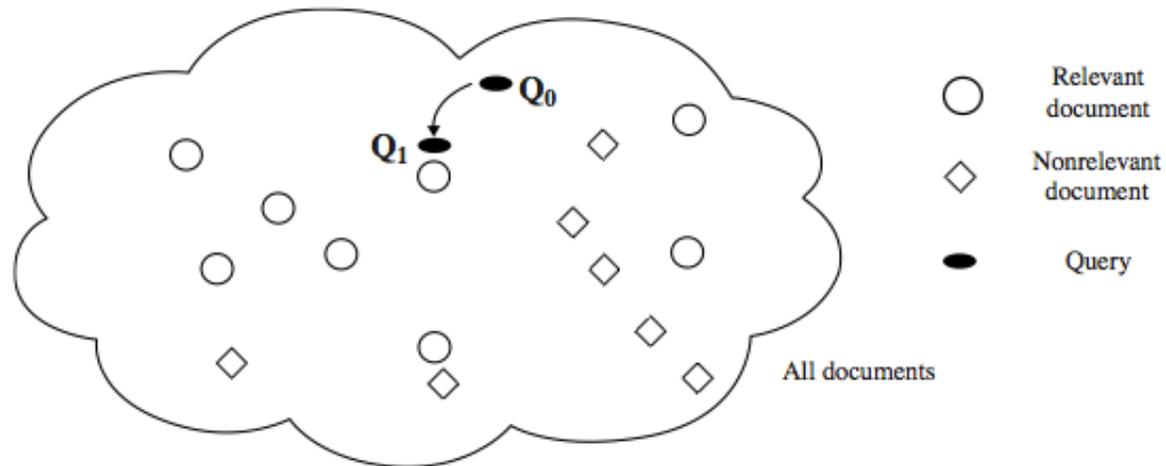


Figure 3.2. Relevance feedback. After feedback, the original query is moved toward the cluster of the relevant documents; see also [Manning et al. \(2008\)](#).

- En ausencia de feedback del usuario, también se suele usar pseudo relevance feedback.

# Representación del Modelo de Usuario III

## Modelo Bayesiano de Syskill & Webert

- El problema es recomendar páginas web de forma personalizada, con un algoritmo que aprenda a medida que el usuario muestra preferencias por ciertos ítems.

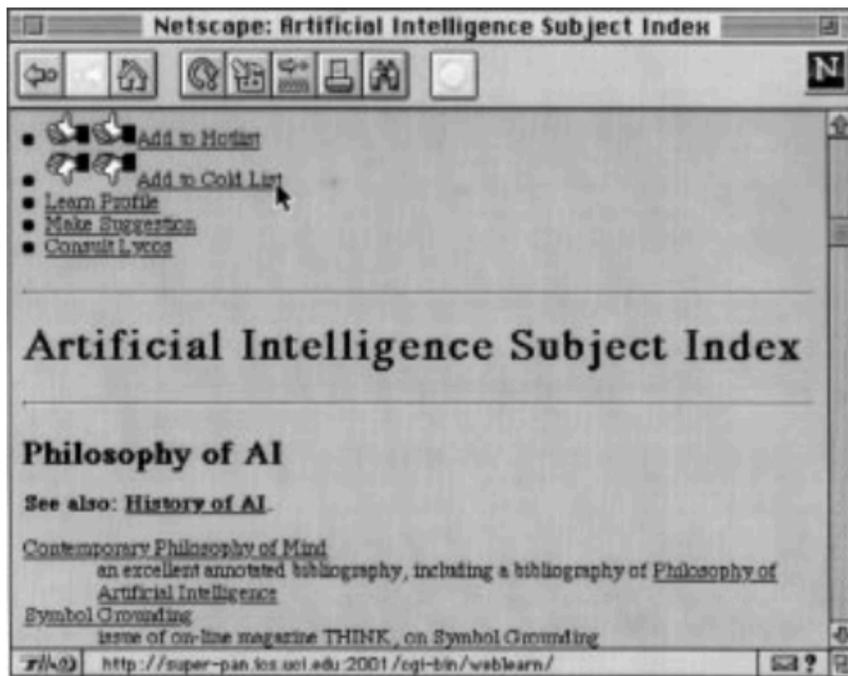


Figure 1. Syskill and Webert interface for rating pages.

- ¿Qué features usar para representar las páginas web y llevar a cabo el aprendizaje?

# Representación del Modelo de Usuario III

## Syskill & Webert: Selección de Features

- No se usaron todas las palabras de la páginas, si no que las que proveían mayor information gain con respecto a clases predefinidas.

$$E(W, S) = I(S) - [P(W = present)I(S_{w=present}) + P(W = absent)I(S_{w=absent})]$$

donde

$$I(S) = \sum_{c \in \{hot, cold\}} -p(S_c) \log_2(p(S_c))$$

- Algunas de las features con mayor info gain.

nirvana	suite	lo
pop	records	rockin
july	jams	songwriting
following	today	vocals
island	tribute	previous
favorite	airplay	noise

Table 1: Some of the words used as features

# Representación del Modelo de Usuario III

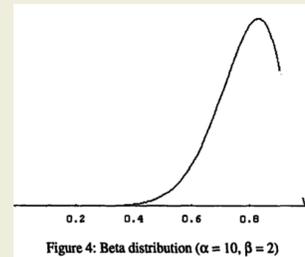
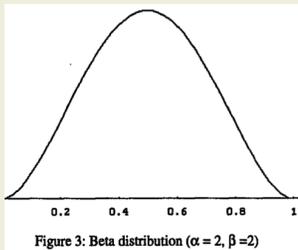
## Modelo Bayesiano de Syskill & Weibert III

- Se usa un clasificador Bayesiano

$$P(C_i | A_1 = V_{1j} \& \dots \& A_n = V_{nj}) \quad \text{donde} \quad P(C_i) \prod_k P(A_k = V_{kj} | C_i)$$

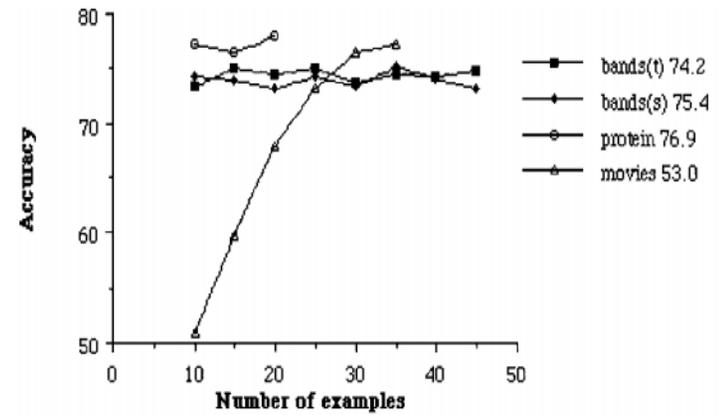
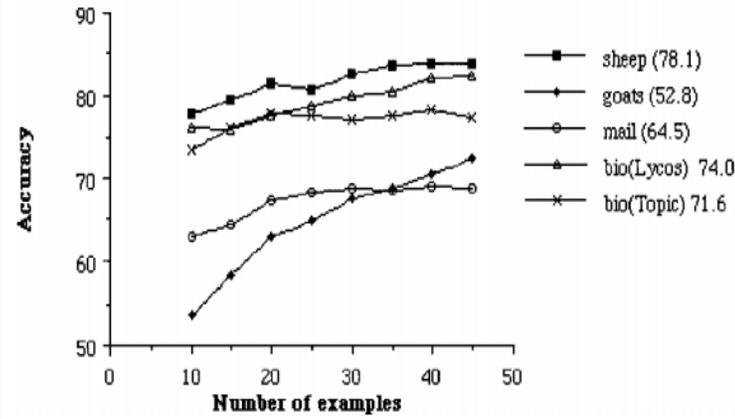
- Y el prior se va actualizando considerando una distribución  $\text{Beta}(\alpha, \beta)$ , con  $\alpha$ : veces que el termino aparece, y  $\beta$ : numero de veces que el termino está ausente.

$$p(\theta | \alpha, \beta) = c \theta^\alpha (1 - \theta)^\beta p(\theta)$$



# Representación del Modelo de Usuario III

Resultados al observar user feedback



# Extensiones

## Tag-Based Recommendation

- D. Parra, P. Brusilovsky. **Improving Collaborative Filtering in Social Tagging Systems for the Recommendation of Scientific Articles.** Web Intelligence 2010, Toronto, Canada
- D. Parra, P. Brusilovsky. **Collaborative Filtering for Social Tagging Systems: an Experiment with CiteULike.** ACM Recsys 2009, New York, NY, USA

# Motivación

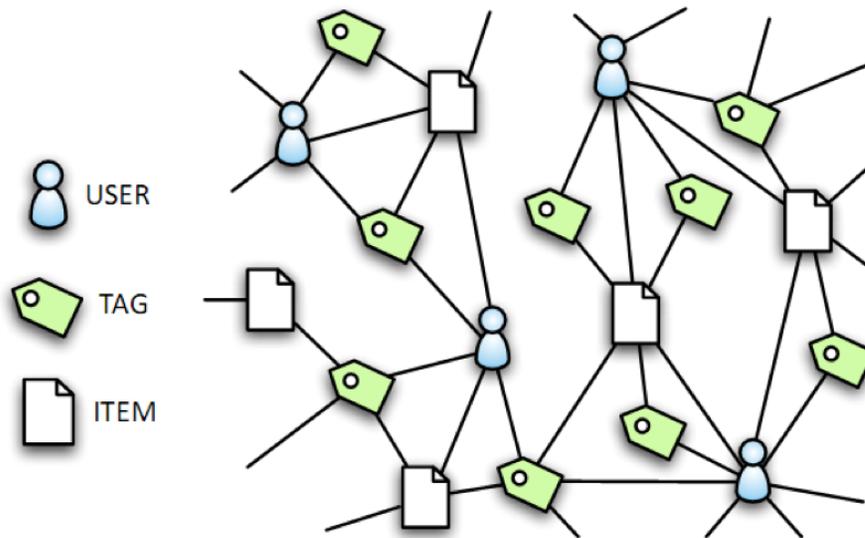
- Es difícil obtener retroalimentación explícito del usuario.

- En Social Tagging Systems el usuario provee "etiquetas" como forma de feedback.

The screenshot shows the CiteULike website interface. The browser window title is "CiteULike: My library 83 articles - Mozilla Firefox". The address bar shows "http://www.citeulike.org/user/denisparra". The page header includes the CiteULike logo, "sponsored by Springer", and a search bar. The main content area is titled "My library 83 articles" and lists several articles. The first article is "Information Retrieval in Folksonomies: Search and Ranking" by Andreas Hotho, Robert Jaschke, Christoph Schmitz, and Gerd Stumme, posted to "tag-recommendation social-bookmarking folkrank" by denisparra on 2010-01-21. The second article is "Tag-LDA for Scalable Real-time Tag Recommendation" by Xiance Si and Maosong Sun, posted to "tag-recommendation lda" by denisparra on 2009-10-28. The third article is "An algorithmic framework for performing collaborative filtering" by Jonathan L. Herlocker, Joseph A. Konstan, Al Borchers, and John Riedl, posted to "evaluation collaborative-filtering" by denisparra on 2009-08-16. The fourth article is "Evaluating collaborative filtering recommender systems" by Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, posted to "evaluation metrics recommender" by denisparra on 2009-08-09. The fifth article is "Being accurate is not enough: how accuracy metrics have hurt recommender systems" by Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl, posted to "evaluation metrics recommender" by denisparra on 2009-08-09. On the right side of the page, there is a section titled "denisparra's tags" which lists all tags in the user's library. The tags are listed with their frequency: recommender (11), collaborative-filtering (11), spreading-activation (7), lecture-3 (7), social-tagging (6), adaptive-hypermedia (6), social-bookmarking (6), lecture-5 (5), tag-based (4), lecture-4 (4), tags (4), folksonomy (4), user-model (4), spread-activation (4), tag-recommendation (4), user-modeling-system (3), clustering (3), lecture-1 (3), evaluation (3), web-20 (2), and case-based (2). The word "Tags" is written in red over the tag list. The word "User" is written in red over the browser window title. The word "Resource" is written in red over the first article title.

# Folksonomy

- Cuando un usuario  $u$  agrega un item  $i$  usando una o más etiquetas  $t_1, \dots, t_n$  se forma un tagging instance.
- La colección de tagging instances produce una folksonomía.



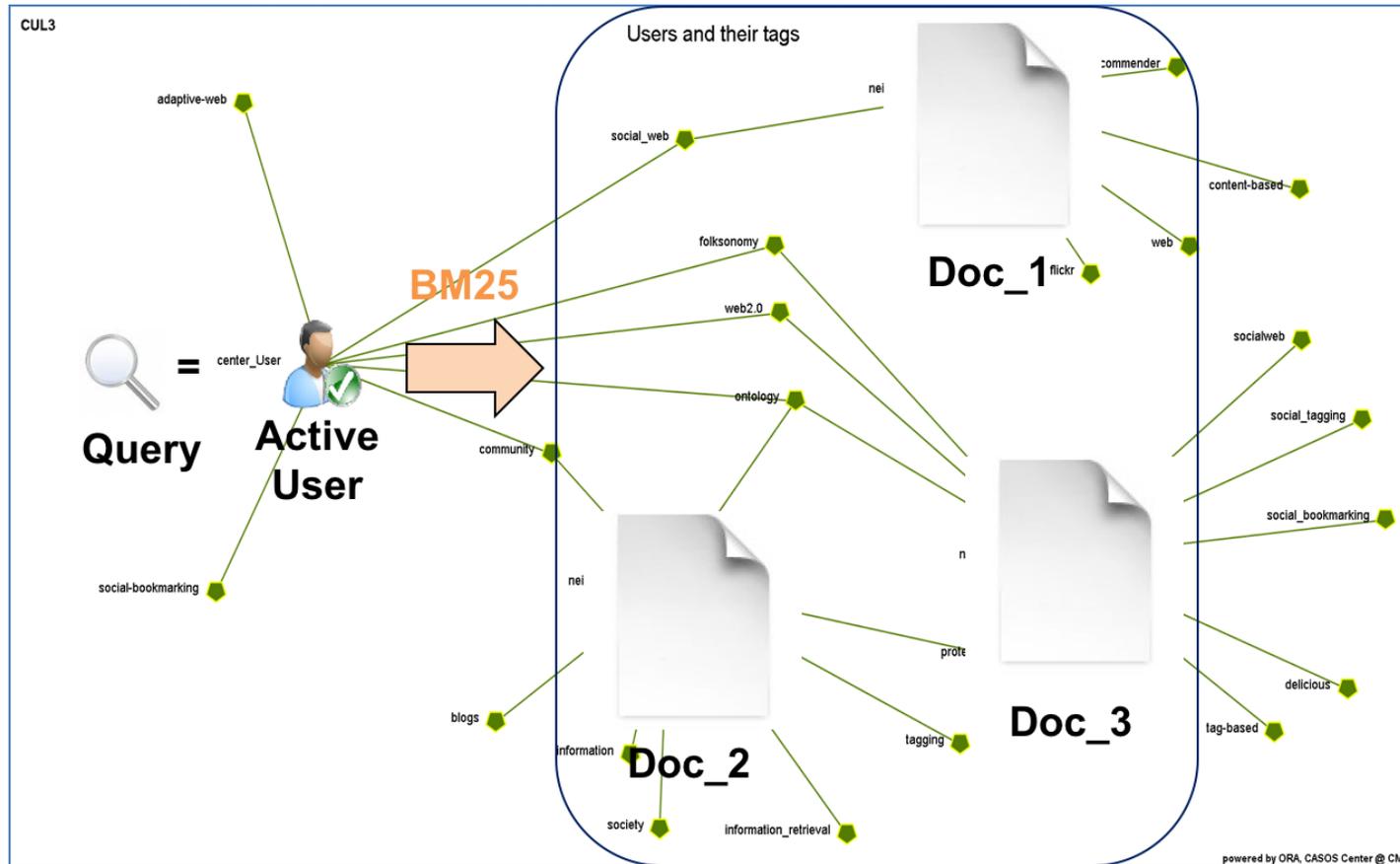
# Usando Filtrado Colaborativo sobre la Folksonomía

Paso 1 : Calcular la similaridad del usuario

TRADITIONAL CF	TAG-BASED CF
Pearson Correlation over ratings	BM25 over social tags

Paso 2 : Incorporar el número de "raters" para rankear items nuevos

# Tag-Based Collaborative Filtering



# Tag-Based Collaborative Filtering

BM25: Obtenemos la similaridad entre usuarios considerando el conjunto de tags de un vecino como un "documento" y los tags del active user como la query.

Usamos Okapi BM25 Retrieval Status Value como medida de similaridad.

$$sim(u, v) = RSV_d = \sum_{i \in q} IDF \cdot \frac{(k_1 + 1)tf_{td}}{k_1((1 - b) + b \times (L_d / L_{ave})) + tf_{td}} \cdot \frac{(k_3 + 1)tf_{tq}}{k_3 + tf_{tq}}$$

Tag frequency in the neighbor (v) profile

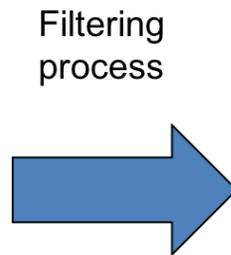
Tag frequency in the active user (u) profile

$$pred'(u, i) = \log_{10}(1 + nbr(i)) \cdot pred(u, i)$$

# Dataset

Crawler para obtener datos desde CiteUlike. 38 días, Junio-Julio 2009

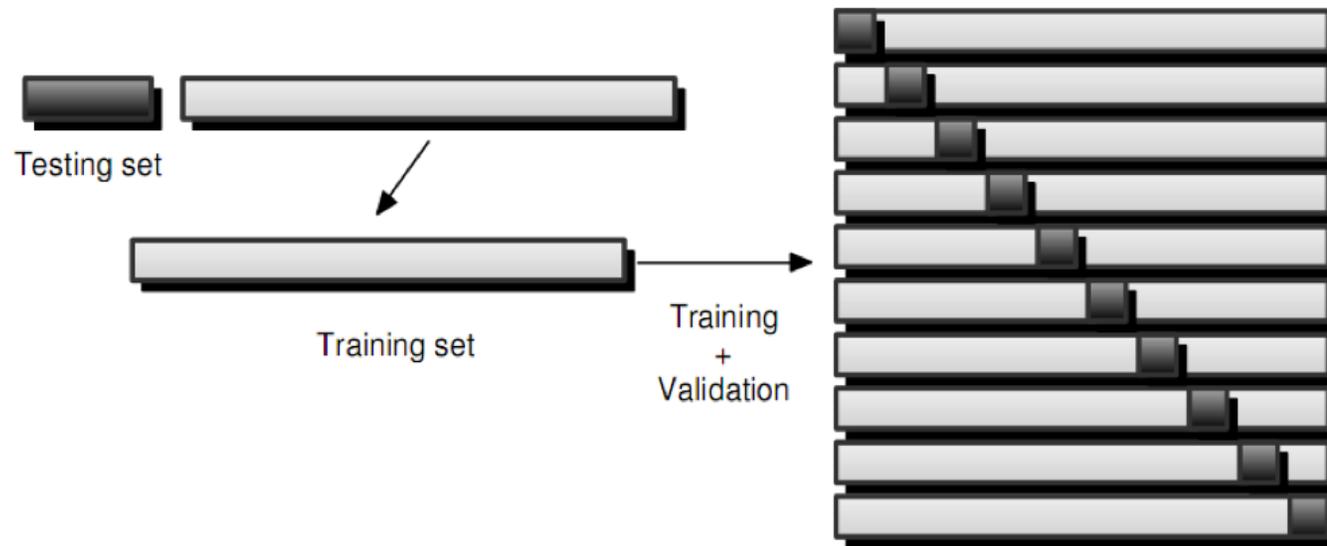
Item	Phase 2 dataset
# users	5,849
# articles	574,907
# tags	139,993
#tagging incidents	2,337,571



Item	# unique instances
# users	784
# items	26,599
# tags	26,009
# posts	71,413
# annotations	218,930
avg # items per user	91
avg # users per item	2.68
avg # tags per user	88.02
avg # users per tag	2.65
avg # tags per item	7.07
avg # items per tag	7.23

# Cross-Validation

- 10-Fold Cross Validation
- Entrenamiento para optimizar parámetro K (vecindario)
- Experimento tomó 12 horas



# Resultados

- BM25 ayuda a obtener más vecinos, con el costo de ruido por aquéllos con pocas etiquetas.

- NwCF ayuda a disminuir el ruido, así es que era natural combinarlos.

	CCF	NwCF	BM25+CCF	BM25+NwCF
MAP@10	0.12875	0.1432*	0.1876**	0.1942***
K (neigh.size)	20	22	21	29
Ucov	81.12%	81.12%	99.23%	99.23%
Significance over the baseline: *p < 0.236, ** p < 0.033, *** p < 0.001				

# Resumen

- Podemos utilizar etiquetas como fuente para encontrar similitud entre los usuarios, como alternativa a algoritmos de recomendación basados en ratings.
- BM25 basado en etiquetas puede reemplazar a la correlación de Pearson para calcular la similitud del usuario en Social Tagging Systems.
- Incorporar el número de usuarios en la fórmula de predicción ayuda a disminuir el ruido por items con muy pocos ratings.

# MM-LDA (Ramage, 2009)

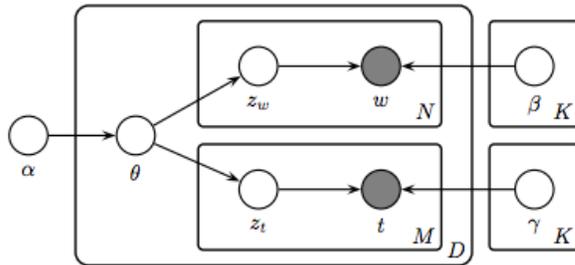


Figure 5: Graphical representation of MM-LDA.

## GENERAL CORPUS

	(MM-)LDA	K-means
Words	0.260	.139
Tags	0.270	.219
Words+Tags	<b>0.307</b>	.225

Figure 7: F-scores for (MM-)LDA and K-means on 13,320 documents. Including tags improves both models significantly versus words alone. MM-LDA (bold) significantly outperforms all other conditions.

## SPECIFIC CORPUS

		(MM-)LDA	K-means
Programming Languages	Words	.288	.189
	Tags	.463	.567
	Words+Tags	.297	.556
Social Sciences	Words	.300	.196
	Tags	.310	.307
	Words+Tags	.302	.308

Figure 10: F-scores for (MM-)LDA and K-means on two representative ODP subtrees. For these tasks, clustering on tags alone can outperform alternatives that use word information.

Ref: Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009, February). Clustering the tagged web. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (pp. 54-63). ACM.

# Combinando Varios Features en un mismo Modelo

Context-Aware Event Recommendation in Event-based Social Networks by Augusto Q. Macedo, Leandro B. Marinho and Rodrygo L. T. Santos

- Recommendations on event-based social networks (EBSNs) often undermines the users' ability to choose the events that best fit their interests.
- The event recommendation problem is intrinsically cold-start. Indeed, events published in EBSNs are typically short-lived and, by definition, are always in the future, having little or no trace of historical attendance.
- THEN: exploit several contextual signals available from EBSNs. In particular, besides content-based signals based on the events' description and collaborative signals derived from users' RSVPs, we exploit social signals based on group memberships, location signals based on the users' geographical preferences, and temporal signals derived from the users' time preferences.

# Referencias

- Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval (Vol. 1, p. 6). Cambridge: Cambridge university press.
- Ramage, D., Heymann, P., Manning, C. D., & Garcia-Molina, H. (2009, February). Clustering the tagged web. In Proceedings of the Second ACM International Conference on Web Search and Data Mining (pp. 54-63). ACM.
- D. Parra, P. Brusilovsky. **Improving Collaborative Filtering in Social Tagging Systems for the Recommendation of Scientific Articles**. Web Intelligence 2010, Toronto, Canada
- D. Parra, P. Brusilovsky. **Collaborative Filtering for Social Tagging Systems: an Experiment with CiteULike**. ACM Recsys 2009, New York, NY, USA