Divide and Transfer: Understanding Latent Factors for Recommendation Tasks

Vidyadhar Rao [1] Rosni K V [1,2] Vineet Padmanabhan [2]

[1] TCS Research Labs, India
vidyadhar.rao@tcs.com

[2] University of Hyderabad, India
rosnikv@gmail.com, vineetcs@uohyd.ernet.in

The 1st Workshop on Intelligent Recommender Systems by Knowledge Transfer & Learning (RecSysKTL) in conjunction with ACM RecSys
Como, Italy, August 27, 2017

Outline

Recommender Systems

We are leaving the age of information and entering the age of recommendation[1]

[1] "The long tail - Why the Future of Business Is Selling Less of More" by Chris Anderson (2006)

Recommender Systems

We are leaving the age of information and entering the age of recommendation[1]

Recommendation Task

- Given a set of users (I), items (J) and a preference matrix R.
- Rating variable $r_{ij} \in \{0, 1\}$ indicates if user i likes item j or not.
- Minimize loss b/w observed and predicted preferences: $min \sum_{i,j}(r_{ij} - \hat{r_{ij}})^2$

Note: $r_{ij} = 0$ can be interpreted in two ways: *either user i is not interested in item j, or user i does not know about item j.*

[1] "The long tail - Why the Future of Business Is Selling Less of More" by Chris Anderson (2006)

Recommender Systems

We are leaving the age of information and entering the age of recommendation[1]

Recommendation Task

- Given a set of users (I), items (J) and a preference matrix R.
- Rating variable $r_{ij} \in \{0, 1\}$ indicates if user i likes item j or not.
- Minimize loss b/w observed and predicted preferences: $min \sum_{i,j}(r_{ij} - \hat{r_{ij}})^2$

Note: $r_{ij} = 0$ can be interpreted in two ways: *either user i is not interested in item j, or user i does not know about item j.*

Two characteristics

- **C**old **S**tart: shortage of information for new users or new items
- **D**ata **S**parsity: users generally rate only a limited number of items.

[1] "The long tail - Why the Future of Business Is Selling Less of More" by Chris Anderson (2006)

Recommender Systems

We are leaving the age of information and entering the age of recommendation[1]

Recommendation Task

- Given a set of users (I), items (J) and a preference matrix R.
- Rating variable $r_{ij} \in \{0, 1\}$ indicates if user i likes item j or not.
- Minimize loss b/w observed and predicted preferences: $min \sum_{i,j}(r_{ij} - \hat{r_{ij}})^2$

Note: $r_{ij} = 0$ can be interpreted in two ways: *either user i is not interested in item j, or user i does not know about item j.*

Two characteristics

- **C**old **S**tart: shortage of information for new users or new items
- **D**ata **S**parsity: users generally rate only a limited number of items.

Factors that influence the recommendations

- Weak algorithmic assumptions about the problem domain: **S**ingle **D**omain vs **C**ross **D**omain
- Intended usage of the model: **P**ast **U**sers vs **N**ew **U**sers
- Type of Recommendations: **U**nknown **I**tems or **N**ew **I**tems vs **D**iverse **I**tems or **S**erendipitous **I**tems

[1] "The long tail - Why the Future of Business Is Selling Less of More" by Chris Anderson (2006)

Content based Filtering

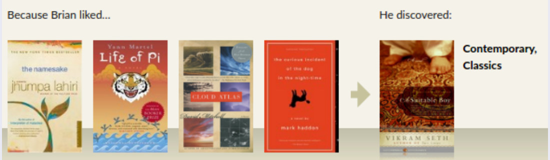
Recommend items that are similar to those that a user liked in the past

Content based Filtering

Recommend items that are similar to those that a user liked in the past

Recommendation based on Topics

- user profiles are created based on the topics that indicate the type of items liked in the past.

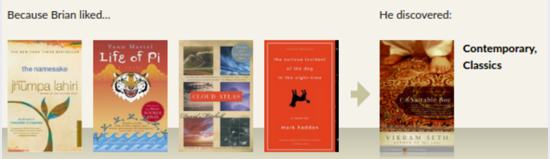


Content based Filtering

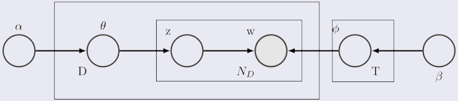
Recommend items that are similar to those that a user liked in the past

Recommendation based on Topics

- user profiles are created based on the topics that indicate the type of items liked in the past.



Latent Dirichlet Allocation [Steyvers and Griffiths, 2007]



- Latent topic space allows to intepret texts in low-dimensional representations.
- documents are drawn from multiple topics: θ
- topics are drawn from a set of words that often appear together: ϕ
- α, β are priors on θ and ϕ, respectively

Collaborative Filtering

Predict interests of a user by collecting preferences from many users

Collaborative Filtering

Predict interests of a user by collecting preferences from many users

Recommendations based on users rating patterns

- Books to read: I read "To kill a Mockingbird" and liked it, you may be interested in it
- Movies to watch: I watched "Okja", you will like it
- Music to listen: "Dunkirk - Supermarine" by Zimmer is todays loved track by most listeners

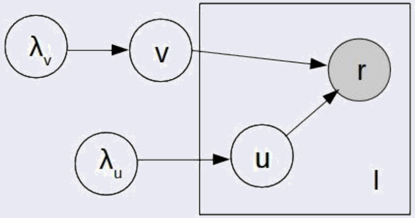
Collaborative Filtering

Predict interests of a user by collecting preferences from many users

Recommendations based on users rating patterns

- Books to read: I read "To kill a Mockingbird" and liked it, you may be interested in it
- Movies to watch: I watched "Okja", you will like it
- Music to listen: "Dunkirk - Supermarine" by Zimmer is todays loved track by most listeners

Probabilistic Matrix Factorization [Mnih and Salakhutdinov, 2008]



- learn common feature space for users and items.

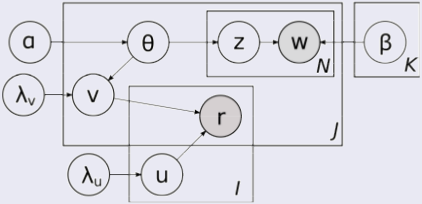
  $$min \sum_{i,j}(r_{ij} - u_i^T v_j)^2 + \lambda_u||u_i||^2 + \lambda_v||v_j||^2$$

- u_i encodes the preferences
- v_j encodes the properties
- λ_v and λ_u are regularization parameters to prevent overfitting
- draw ratings from Guassian distribution: $\hat{r_{ij}} \sim \mathcal{N}(u_i^T v_j, c_{ij})$, c_{ij} is a confidence parameter.

Collaborative Topic Regression

Recommend based on both content and user preferences[1]

Collaborative Topic Regression [Wang and Blei, 2011]



- Do Latent Dirichlet Allocation to estimate θ
- With θ we can maximize the log-likelihood,

$$\mathcal{L} = \log P(U, V, R|\theta, \lambda_v, \lambda_u, \alpha, \beta) = C - \sum_{i,j} \frac{c_{ij}}{2}(r_{ij} - u_i^T v_j)^2 - \frac{\lambda_u}{2} \sum_i ||u_i||^2 - \frac{\lambda_v}{2} \sum_j ||v_j - \theta_j||^2$$

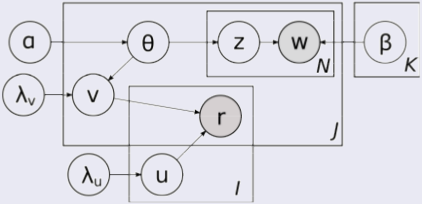
- latent variable ϵ_j offsets the topic proportion θ_j.

$$v_j = \theta_j + \epsilon_j$$

$$E\left[r_{ij} \mid u_i, \theta_j, \epsilon_j\right] = u_i^T(\theta_j + \epsilon_j)$$

[1]https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/

Collaborative Topic Regression [Wang and Blei, 2011]



- Do Latent Dirichlet Allocation to estimate θ
- With θ we can maximize the log-likelihood,

$$\mathcal{L} = \log P(U, V, R|\theta, \lambda_v, \lambda_u, \alpha, \beta) = C - \sum_{i,j} \frac{c_{ij}}{2}(r_{ij} - u_i^T v_j)^2 - \frac{\lambda_u}{2} \sum_i ||u_i||^2 - \frac{\lambda_v}{2} \sum_j ||v_j - \theta_j||^2$$

- latent variable ϵ_j offsets the topic proportion θ_j.

$$v_j = \theta_j + \epsilon_j$$
$$E\left[r_{ij} \mid u_i, \theta_j, \epsilon_j\right] = u_i^T(\theta_j + \epsilon_j)$$

Benefits over traditional models

- generalization to unseen or new items. $\epsilon_j = 0$
- generate interpretable user profiles using topics.

[1]https://open.blogs.nytimes.com/2015/08/11/building-the-next-new-york-times-recommendation-engine/

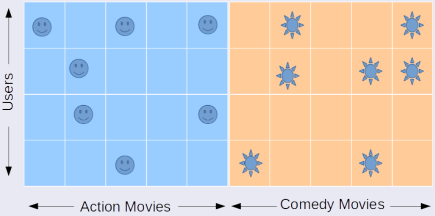
Weak Algorithmic Assumptions

Desire to produce meaningful recommendations that align with the target environment

Single-domain recommendations: Need for hybrid models

Method	aggregate user preferences	recommend new or unseen items	interpretable user profiles
Collaborative Filtering (PMF)	✓	×	×
Content based Recommendations (LDA)	×	✓	✓
Collaborative Topic Regression (CTRlda)	✓	✓	✓

Cross-domain recommendations: Same users might show different rating preferences



- latent features derived from content are naturally different across domains.
- naive feature fusion might generate latent features that may degrade quality of recommendations.
- understanding the semantic aspects of the latent factors is highly desirable in cross domain research.

Semantic Representation of Words

Understand the semantic aspects of the topical space

Latent Semantic Space

- background information: words occuring in many documents and distributed over many topics in LDA
- document-specific information: words occuring in a few documents and distributed over a certain topics in LDA
- general topics: explains the documents with compact representation of topics in LDA.

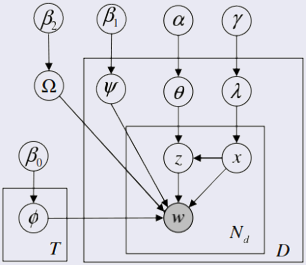
Semantic Representation of Words

Understand the semantic aspects of the topical space

Latent Semantic Space

- background information: words occuring in many documents and distributed over many topics in LDA
- document-specific information: words occuring in a few documents and distributed over a certain topics in LDA
- general topics: explains the documents with compact representation of topics in LDA.

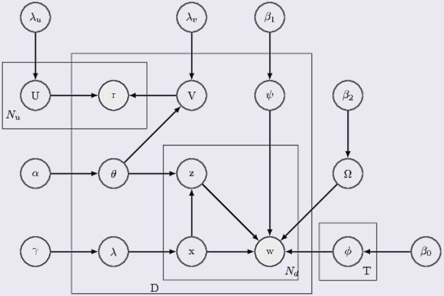
Special Word Topic model [Chemudugunta et al., 2007]



- accounts for both general and specific aspects of documents.
- each document is represented as a combination of
  - background distribution over common words: Ω
  - mixture distribution over general topics: θ
  - distribution of words that are treated as specific to that document: ψ
- SWB introduces additional switch variables into the LDA model to account for multiple word distributions
- $\alpha, \beta_0, \beta_1, \beta_2, \gamma$ - all weak symmetric priors.

Recommendations do not depend on document-specific and corpus-specific information

Combining PMF and SWB model to derive latent factors

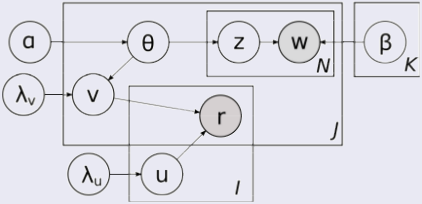


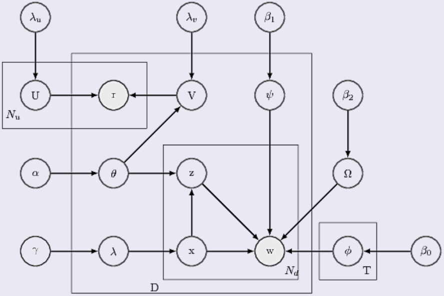
Select a background distribution over words $\Omega|\beta_2 \sim Dir(\beta_2)$
for *each topic* $k \in 1 \ldots T$ **do**
  Select a word distribution $\phi_k|\beta_0 \sim Dir(\beta_0)$
end
for *each document* $d \in 1 \ldots D$ **do**
  Select a distribution over topics $\theta_d|\alpha \sim Dir(\alpha)$
  Select a special-words distribution over words $\psi_d|\beta_1 \sim Dir(\beta_1)$
  Select a distribution over switch variables $\lambda_d|\gamma \sim Beta(\gamma)$
  for $n = 1 : N_d$ *words in document* d **do**
    Select a switch variable $x_{dn}|\lambda_d \sim Mult(\lambda_d)$
    Select $z_{dn}|\{\theta_d, x_{dn}\} \sim Mult(\theta_d)\delta(z_{dn}, SW)\delta(z_{dn}, BG)$
    Generate a word: $w_{dn}|\{z_{dn}, x_{dn}, \phi, \psi_d, \Omega\} \sim Mult(\phi_{z_{dn}})Mult(\psi_d)Mult(\Omega)$
  end
end
for *user* $i \in 1 \ldots N_u$ **do**
  Draw $u_i \sim \mathcal{N}(0, \lambda_u^{-1}\mathcal{I}_T)$
end
for *item* $j \in 1 \ldots D$ **do**
  Draw $\epsilon_j \sim \mathcal{N}(0, \lambda_v^{-1}\mathcal{I}_T)$
  Compute $v_j = \epsilon_j + \theta_j$
end
for *user-item pair* $(i, j) \in \{1 \ldots N_u\} \times \{1 \ldots D\}$ **do**
  Draw $r_{ij} \sim \mathcal{N}(u_i^T v_j, c_{ij})$
end

CTR Framework

Contrasting latent factors derived in CTRlda and CTRswb

Graphical Models of CTRlda and CTRswb





- CTRlda uses topic proportions of LDA model i.e, $CTRlda = PMF \circ LDA$
- CTRswb uses general topic proportions of SWB model i.e, $CTRswb = PMF \circ SWB$
- For learning, and inference CTRswb does no hard splitting or removal of topics.

Datasets

CiteULike Dataset[1]

- contains 204, 986 pairs of observed ratings with 5, 551 users and 16, 980 articles, and with 99.7825% sparsity.
- each user has 37 articles in their library on an average and only 7% of the users has more than 100 articles.
- title and abstract of articles constitute item descriptions.
- **S**ingle **D**omain **R**ecSys: recommend new or unseen articles to a scholar

MovieLens Dataset[2]

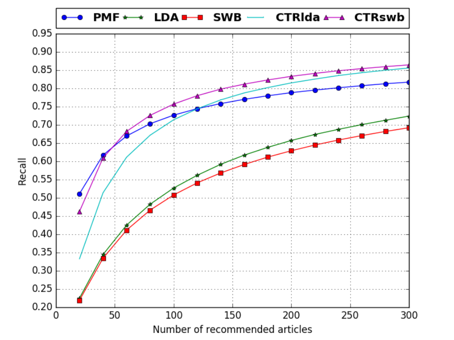
- contains 1 million user-movie ratings across 19 genres.
- we use five genres with most ratings: Action, Comedy, Drama, Romance, Thriller.
- movie descriptions are crawled from IMDb[3]
- **C**ross **D**omain **R**ecSys: recommend movies in a target genre based on ratings from 4 other genres.

Movie Genre	No.Items	No.Users	No.Ratings	Rating Ratio
Drama	1,493	5,881	352,834	0.040
Comedy	1,163	5,881	354,455	0.052
Thriller	485	5,881	188,968	0.066
Romance	459	5,881	146,916	0.054
Action	495	5,881	256,515	0.088
Total	4,095	5,881	1,299,688	0.054

[1]http://www.citeulike.org/
[2]http://grouplens.org/datasets/movielens/
[3]http://www.imdb.com

Single Domain recommendations

Recall measure - CiteuLike Dataset

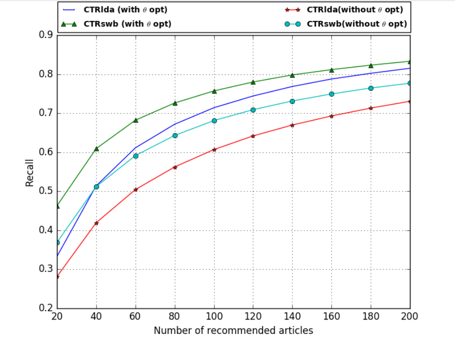


$$recall@K = \frac{\text{No. of items correctly recommended}}{\text{Total number of items the user likes}}$$

- pure latent factor models and pure content models underperform w.r.t CTR models.
- CTRswb consistently outperforms CTRlda.
- margin of improvement for smaller number of recommendations is large.

Single Domain recommendations

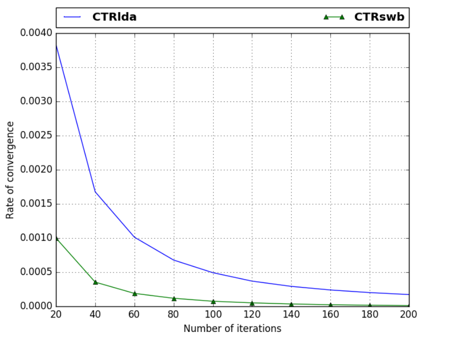
θ optimization - CiteULike Dataset



- with or without θ optimization, CTRswb shows superior performance over CTRlda.
- CTRswb could compute θ in a precise manner.
- automatically discards non-useful topic proportions while learning latent features.

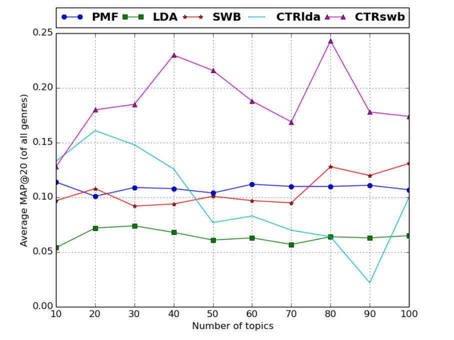
Single Domain recommendations

Convergence curve - CiteULike Dataset



- Convergence curve of CTRlda and CTRswb during θ optimization.
- $covergenceRate = \frac{\mathcal{L} - \mathcal{L}_{old}}{\mathcal{L}_{old}}$ where $\mathcal{L} = \log P(U, V, R|\theta, \lambda_v, \lambda_u, \alpha, \beta)$
- Faster convergence of CTRswb compared to CTRlda.
- For recommendations, θ proportions from SWB model are effective compared to LDA model.

Cold-start scenario - MovieLens Dataset



- **C**old **S**tart scenario and the impact of number of topics on the recommendations
- illustrates potential problems with the learned topics obtained from feature fusion of multiple domains.
- CTRswb explicitly models these aspects and shows improved latent features space over others.

Cross Domain recommendations

Cold-start scenario - MovieLens Dataset

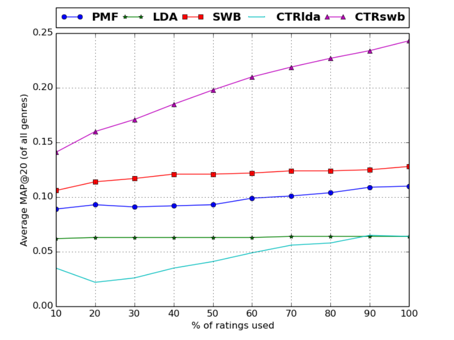
Comparison of different models with 80 latent factors for various **C**old **S**tart scenarios.

Target Genre	Method	MAP@20	P@10	P@20	R@10	R@20
Action	PMF	0.133	0.072	0.069	0.013	0.024
	LDA	0.057	0.025	0.026	0.005	0.01
	SWB	0.244	0.136	0.11	0.035	0.052
	CTRlda	0.099	0.061	0.057	0.013	0.025
	CTRswb	**0.306**	**0.176**	**0.14**	**0.051**	**0.07**
Drama	PMF	0.09	0.039	0.038	0.006	0.012
	LDA	0.075	0.027	0.027	0.004	0.009
	SWB	0.1	0.044	0.041	0.009	0.016
	CTRlda	0.024	0.011	0.013	0.001	0.004
	CTRswb	**0.235**	**0.07**	**0.055**	**0.02**	**0.026**
Romance	PMF	0.099	0.048	0.046	0.015	0.028
	LDA	0.038	0.012	0.014	0.004	0.009
	SWB	0.094	0.029	0.025	0.024	0.037
	CTRlda	0.056	0.036	0.024	0.022	0.027
	CTRswb	**0.367**	**0.084**	**0.06**	**0.061**	**0.07**

- CTRlda breaks down when latent factors are learned from feature fusion from multiple domains
- CTRswb model significantly improves over CTRlda and other methods in all the cold-start scenarios

Cross Domain recommendations

Sparse data scenario - MovieLens Dataset



- **D**ata **S**parse scenatio and the impact of number of ratings from source domains.
- CTRlda does not scale well under sparse user ratings.
- CTRswb shows better performance, by large margin, than others.
- CTRswb robust even under extreme sparse data scenarios.

Summary

Take away

Method	user preferences	new or unseen items	interpretable user profiles	semantic aspects	problem domain
Collaborative Filtering - PMF	✓	×	×	×	single
Content based Recommendations - LDA/SWB	×	✓	✓	×	single
Collaborative Topic Regression - CTRlda	✓	✓	✓	×	single
Divide and Transfer Topics - CTRswb (ours)	✓	✓	✓	✓	single, cross

Our claims

- Understanding the latent factors could give a hint on how to transfer useful information.
- Semantic aspects of latent factors is highly desirable in cross domain research.
- Divide and Transfer Approach - CTRswb
    - explicitly learns latent features by understanding the multiple topic proportions.
    - improves latent feature space even under extreme cold-start and data sparse situations.
    - effective and robust in both single and cross-domain recommendations.

Summary

Take away

Method	user preferences	new or unseen items	interpretable user profiles	semantic aspects	problem domain
Collaborative Filtering - PMF	✓	×	×	×	single
Content based Recommendations - LDA/SWB	×	✓	✓	×	single
Collaborative Topic Regression - CTRlda	✓	✓	✓	×	single
Divide and Transfer Topics - CTRswb (ours)	✓	✓	✓	✓	single, cross

Our claims

- Understanding the latent factors could give a hint on how to transfer useful information.
- Semantic aspects of latent factors is highly desirable in cross domain research.
- Divide and Transfer Approach - CTRswb
    - explicitly learns latent features by understanding the multiple topic proportions.
    - improves latent feature space even under extreme cold-start and data sparse situations.
    - effective and robust in both single and cross-domain recommendations.

Outlook

- Explore cross-domain recommendation scenarios in heterogeneous settings (e.g movies to books).
- Extend to user ratings on a scale from multiple domains.

References

Chemudugunta, C., Smyth, P., and Steyvers, M. (2007).
Modeling general and specific aspects of documents with a probabilistic topic model.
In *NIPS*.

Mnih, A. and Salakhutdinov, R. R. (2008).
Probabilistic matrix factorization.
In *NIPS*.

Steyvers, M. and Griffiths, T. (2007).
Probabilistic topic models.
Handbook of latent semantic analysis.

Wang, C. and Blei, D. M. (2011).
Collaborative topic modeling for recommending scientific articles.
In *ACM SIGKDD*.

Thank you for your attention!

Back up Slides: Cold-start scenario - MovieLens Dataset

Comparison of different models with 80 latent factors for various cold start scenarios.

Target Genre	Method	MAP@20	P@10	P@20	R@10	R@20
Comedy	PMF	0.101	0.05	0.049	0.008	0.014
	LDA	0.073	0.024	0.027	0.004	0.009
	SWB	0.122	0.059	0.05	0.009	0.016
	CTRlda	0.059	0.029	0.026	0.007	0.012
	CTRswb	**0.147**	**0.074**	**0.061**	**0.011**	**0.018**
Thriller	PMF	0.127	0.063	0.06	0.016	0.029
	LDA	0.076	0.035	0.028	0.012	0.018
	SWB	0.079	0.041	0.04	0.01	0.02
	CTRlda	0.084	0.038	0.031	0.016	0.027
	CTRswb	**0.162**	**0.09**	**0.073**	**0.022**	**0.034**

Single Domain recommendations

Back up Slides: Qualitative results - CiteULike Dataset

	CTR(with LDA)	
	User 1	In User's lib?
Top 3 topics	Topic 1 : information seeking sources overload retrieval encountered relating opposed fact unfortunately concerned	
	Topic 2 : developed widely requires challenge list lists typical aggregation allows combining rapid	
	Topic 3 : aspects account working serve showing two event aspect describe first provides	
	Doc Id : Doc Title	
Top 10 articles	7177 : Information Retrieval, 2nd edition	N
	6347 : Folksonomies: Tidying up Tags?	Y
	7545 : Usage patterns of collaborative tagging systems	Y
	3981 : The Structure of Collaborative Tagging Systems	Y
	4020 : Information Retrieval.	N
	5034 : Folksonomy as a Complex Network	N
	7922 : Collaborative Tagging and Semiotic Dynamics	Y
	615 : Folksonomies - Cooperative Classification and Communication Through Shared Metadata	Y
	7135 : Collaborative Tagging as a Knowledge Organisation and Resource Discovery Tool	Y
	3096 : Shirky: Ontology is Overrated -- Categories, Links, and Tags	Y

Back up Slides: Qualitative results - CiteULike Dataset

	CTR(with SWB)	
	User 1	In user's lib?
Top 3 topics	Topic 1 : individual common ability core activities shared share activity explored awareness joint	
	Topic 2 : natural classification categories subject category hand taxonomy viewed schemes categorization classified	
	Topic 3 : information seeking mutual overload pim needs pieces encountered shannon professionals provision	
	Doc Id : Doc Title	
Top 10 articles	3981 : The Structure of Collaborative Tagging Systems	Y
	3096 : Shirky: Ontology is Overrated -- Categories, Links, and Tags	Y
	615 : Folksonomies - Cooperative Classification and Communication Through Shared Metadata	Y
	9103 : HT06, tagging paper, taxonomy, Flickr, academic article, to read	N
	7922 : Collaborative Tagging and Semiotic Dynamics	Y
	7545 : Usage patterns of collaborative tagging systems	Y
	6347 : Folksonomies: Tidying up Tags?	Y
	8158 : Why do tagging systems work?	Y
	2350 : Social Bookmarking Tools (I): A General Review	Y
	8160 : Visualizing Tags over Time	N