Text Retrieval Methods for Item Ranking in Collaborative Filtering

Alejandro Bellogín[1], Jun Wang[2], and Pablo Castells[1]

[1]Escuela Politécnica Superior, Universidad Autónoma de Madrid
[2]Department of Computer Science, University College London
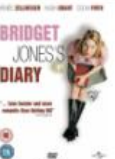
@abellogin
alejandro.bellogin@uam.es

Collaborative Filtering – in a glimpse

- Input: a rating matrix

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- The users

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- The items

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- A rating: from a user towards an item

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- User profile

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- Item profile

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆
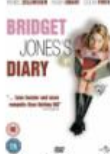
Collaborative Filtering – in a glimpse

- Unknown rating

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★★★	★★★★★	★★★★★			★★★★★		
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

Collaborative Filtering – in a glimpse

- Goal

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris	★★★☆☆	★★★★☆	★★★★★			★☆☆☆☆		?
Dave		★★★★★	★★★★★	★★★★★				★☆☆☆☆
Will		★★☆☆☆			★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★	★★★★☆	★★★★☆				★★☆☆☆

If user Boris watched *Love Actually*, how would he rate it?

Collaborative Filtering – in a glimpse

- Prediction: how Boris rated similar items

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris		★★★★☆	★★★★★		★★★★☆	★☆☆☆☆	★★★☆☆	?
Dave		★★★★★	★★★★★	★★★★★	★★☆☆☆			★☆☆☆☆
Will					★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★		★★★★☆		★★★☆☆	★★★★☆	★★☆☆☆

Text Retrieval

Collaborative Filtering?
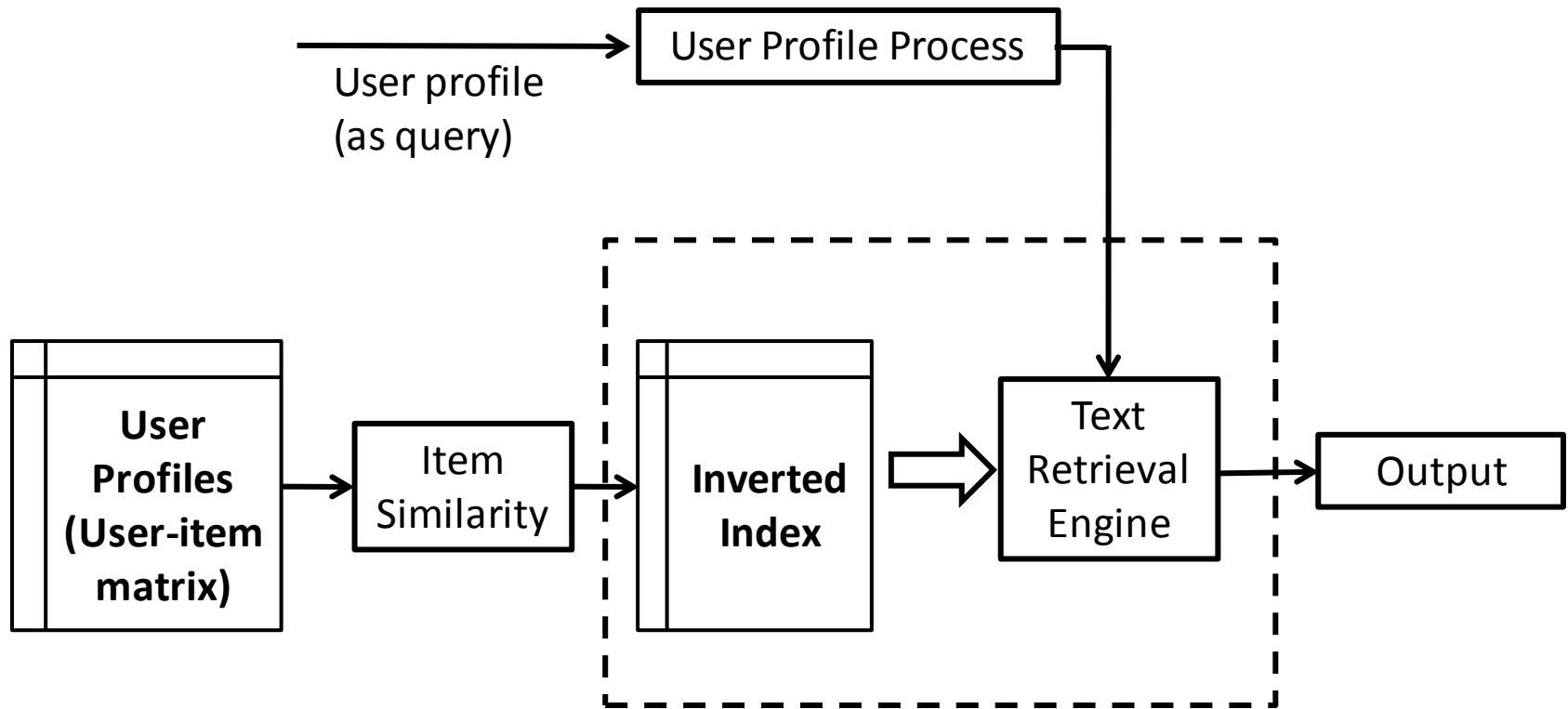
In this work

- A first attempt
  - Item ranking
  - Item-based CF
- Good results
- Improvements – current work
  - More models
  - Rating prediction

- Now… algorithmic details

Escuela
Politécnica
Superior

Text Retrieval

- In (Metzler & Zaragoza, 2009)

$$s(q,d) = \sum_{t \in g(q)} s(q,d,t)$$

- In particular: *factored form*

$$s(q,d,t) = w_1(q,t)\, w_2(d,t)$$

UNIVERSIDAD AUTONOMA
DE MADRID

Escuela
Politécnica
Superior

Text Retrieval

- Examples
  - TF:

$$w_1(q,t) = \text{qf}(t)$$

$$w_2(d,t) = \text{tf}(t,d)$$

  - TF-IDF:

$$w_1(q,t) = \text{qf}(t)$$

$$w_2(d,t) = \text{tf}(t,d)\log\left(\frac{N}{\text{df}(t)}\right)$$

  - BM25:

$$w(q,t)_1 = \frac{(k_3+1)\text{qf}(t)}{k_3+\text{qf}(t)}$$

$$w(d,t)_2 = \log\left(\frac{N-\text{df}(t)+0.5}{\text{df}(t)+0.5}\right)\frac{(k_1+1)\text{tf}(t,d)}{k_1\left((1-b)+b\cdot\text{dl}(d)/\overline{\text{dl}}\right)+\text{tf}(t,d)}$$

Collaborative Filtering

- Standard item-based formulation (Adomavicius & Tuzhilin 2005)

$$\mathrm{rat}(u,i) = \sum_{j \in I_u} \frac{\mathrm{sim}(i,j)}{\sum_{j \in I_u} \mathrm{sim}(i,j)} \mathrm{rat}(u,j)$$

Collaborative Filtering

- Standard item-based formulation (Adomavicius & Tuzhilin 2005)

$$\mathrm{rat}(u,i) = \sum_{j \in I_u} \frac{\mathrm{sim}(i,j)}{\sum_{j \in I_u} \mathrm{sim}(i,j)} \mathrm{rat}(u,j)$$

	Die Hard	Mission: Impossible	GoldenEye	Casino Royale	Titanic	Notting Hill	Bridget Jones's Diary	Love Actually
Boris		★★★★☆	★★★★★		★★★★☆	★☆☆☆☆	★★★☆☆	?
Dave		★★★★★	★★★★★	★★★★★	★★☆☆☆			★☆☆☆☆
Will					★★★★★	★★★★★	★★★☆☆	★★★★☆
George	★★★★☆	★★★★★		★★★★☆		★★★☆☆	★★★★☆	★★☆☆☆

UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

Collaborative Filtering

- Standard item-based formulation

$$\mathrm{rat}(u,i) = \sum_{j \in I_u} \frac{\mathrm{sim}(i,j)}{\displaystyle\sum_{j \in I_u} \mathrm{sim}(i,j)} \mathrm{rat}(u,j)$$

- More general

$$\mathrm{rat}(u,i) = \sum_{j \in g(u)} f(u,i,j) = \sum_{j \in g(u)} f_1(u,j) f_2(i,j)$$

UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

Text Retrieval for Collaborative Filtering

- In item-based Collaborative Filtering

$$\text{tf}(t, d) = \text{sim}(i, j)$$
$$\text{qf}(t) = \text{rat}(u, j)$$

$$t \approx j$$
$$d \approx i$$
$$q \approx u$$
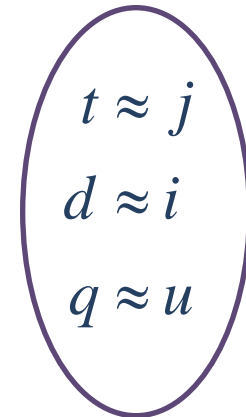
- Apply different models
  - With different normalizations and norms: s_{qd}, L_1 and L_2

s_{qd}		Document	
		No norm	Norm (/|D|)
Query	No norm	s_{00}	s_{01}
	Norm (/|Q|)	s_{10}	s_{11}

Text Retrieval for Collaborative Filtering

- TF L1 s01 equivalent to item-based CF

$$s(q,d) = \sum_{t \in g(q)} w_1(q,t) w_2(d,t) = \sum_{t \in g(q)} \mathrm{qf}(t) \frac{\mathrm{tf}(t,d)}{\sum_{t \in g(q)} \mathrm{tf}(t,d)}$$

$$\mathrm{rat}(u,i) = \sum_{j \in I_u} \mathrm{rat}(u,j) \frac{\mathrm{sim}(i,j)}{\sum_{j \in I_u} \mathrm{sim}(i,j)}$$

$$\mathrm{tf}(t,d) = \mathrm{sim}(i,j)$$
$$\mathrm{qf}(t) = \mathrm{rat}(u,j)$$

UNIVERSIDAD AUTONOMA DE MADRID

Escuela Politécnica Superior

Results

- ## Movielens 1M
  - Movielens100k: equivalent results
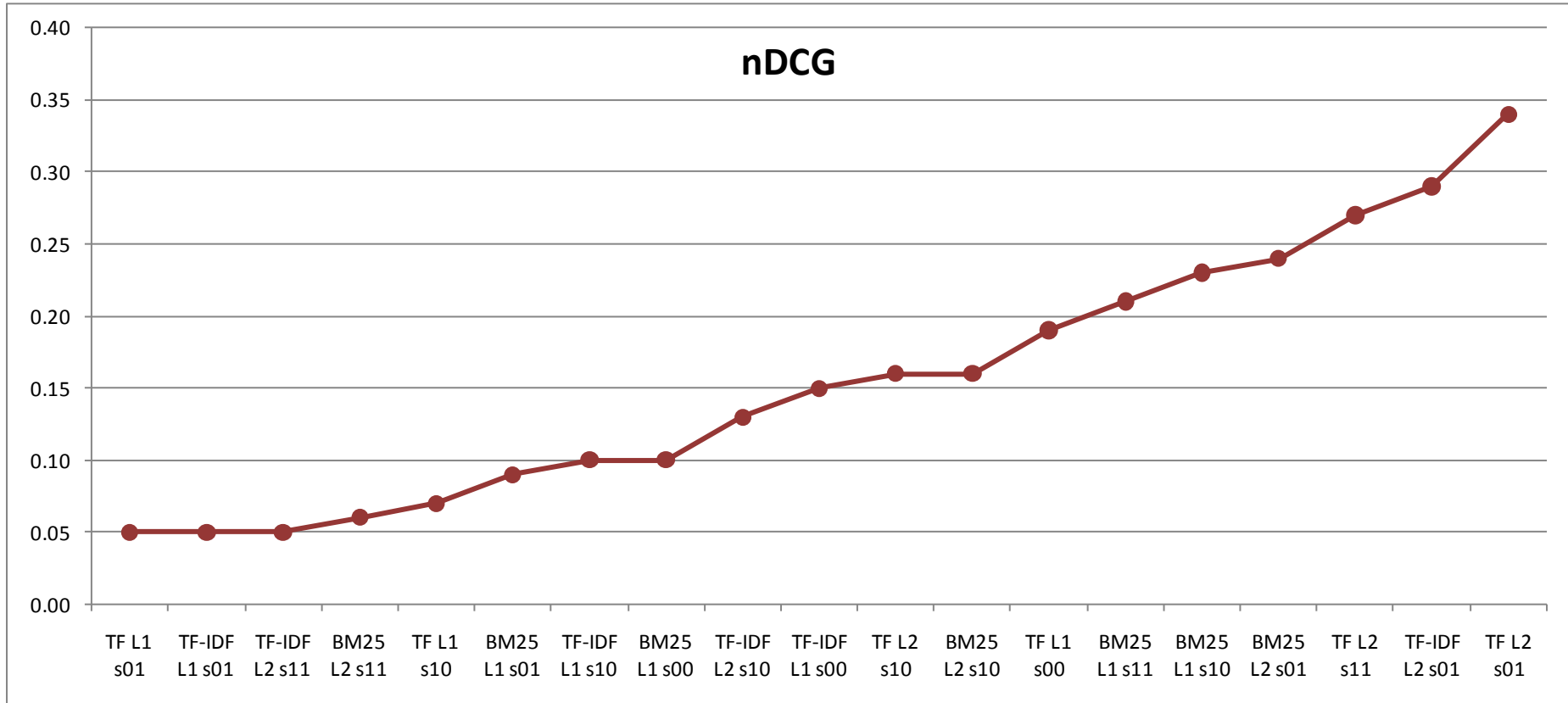
- **TF L1 s01 equivalent to item-based CF (baseline)**

Conclusions

- It is possible to use Text Retrieval methods in rating-based Collaborative Filtering

- Our methods outperform classic Collaborative Filtering methods

- …Questions?

References

- Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: a survey of the state-of-the-art and possible extensions. IEEE TKDE 17(6), 734-749 (2005)

- Metlzer, D., Zaragoza, H.: Semi-parametric and non-parametric term weighting for information retrieval. LNCS, vol. 5766, pp. 42-53. Springer (2009)