Predicting Neighbor Goodness in Collaborative Filtering

Alejandro Bellogín and Pablo Castells

{**alejandro.bellogin**, pablo.castells}@uam.es

Universidad Autónoma de Madrid
Escuela Politécnica Superior

Introduction: Recommender Systems

- Recommender Systems (RS) goal

objects

users

	i_1		i_k		i_m
u_1	r_{11}		r_{1k}		r_{1m}
u_j	r_{j1}		?		r_{jm}
u_n	r_{n1}		r_{nk}		r_{nm}

how is predicted the value of r_{jk}?

Introduction: Recommender Systems

- Collaborative filtering (CF) based on users

items

users

	i_1		i_k		i_m
u_1	r_{11}		r_{1k}		r_{1m}
u_j	r_{j1}		?		r_{jm}
u_n	r_{n1}		r_{nk}		r_{nm}

- Producing recommendations in UBCF:

$$g(u_m, i_n) = \frac{\sum_{u_j \in N[u_m]} \mathrm{sim}(u_m, u_j) \times r_{j,n}}{\sum_{u_j \in N[u_m]} \left|\mathrm{sim}(u_m, u_j)\right|}$$

Introduction: Recommender Systems

- Collaborative filtering (CF) based on items

items

users

	i_1		i_k		i_m
u_1	r_{11}		r_{1k}		r_{1m}
u_j	r_{j1}		?		r_{jm}
u_n	r_{n1}		r_{nk}		r_{nm}

- Producing recommendations in IBCF:

$$g\left(u_m, i_n\right) = \frac{\sum_{i_j \in N[i_n]} \mathrm{sim}\left(i_n, i_j\right) \times r_{m,j}}{\sum_{i_j \in N[i_n]} \left|\mathrm{sim}\left(i_n, i_j\right)\right|}$$

Motivation

- Proliferation and variety of input information in IR systems
- Performance prediction in IR: adjust retrieval strategies according to the value of the prediction function
  - In classic retrieval: query effectiveness
  - Applications: query expansion, meta-search, distributed IR, etc.
- In CF, each neighbor can be seen as a different source of information
  - ¿with different weight? (besides the similarity value)

Performance prediction in IR

- Mostly addressed: query performance prediction in ad-hoc retrieval

- Approaches (Hauff et al. 2008)

  - Pre-retrieval
    - Pros: The prediction can be taken into account to improve the retrieval process itself
    - Cons: Effectiveness cues available after the retrieval are not exploited

  - Post-retrieval
    - Pros: Better prediction accuracy
    - Cons: Problem with computational efficiency

Clarity score in ad-hoc retrieval

- Distance (relative entropy) between query and collection language models (Cronen-Townsend et al. 2002)

$$\text{clarity}(q) = \sum_{w \in V} P(w|q) \log_2 \frac{P(w|q)}{P_{coll}(w)}$$

$$P(w|q) = \sum_{d \in R} P(w|d) P(d|q), \quad P(q|d) = \prod_{w_q \in q} P(w_q|d)$$

$$P(w|d) = \lambda P_{ml}(w|d) + (1-\lambda) P_{coll}(w)$$

- It captures the (lack of) ambiguity in a query with respect to the collection
  - Queries whose likely relevant documents are a mix of disparate topics receive a lower score than those with a topically-coherent result set.
  - Strong correlation between this score and the performance (average precision) of the result set

Performance prediction in Recommender Systems

- Rating prediction as dynamic aggregation

  - Each user's neighbor can be seen as a retrieval subsystem whose output is to be combined to form the final system output
  - Utility of an item for a user:

$$r(u,i) = \bar{r}(u) + C \sum_{v \in \mathrm{N}[u]} \mathrm{sim}(u,v) \cdot \left(r(v,i) - \bar{r}(v) \right)$$

- Using dynamic aggregation:

$$r(u,i) = \bar{r}(u) + C \sum_{v \in \mathrm{N}[u]} \gamma(v,u,i) \cdot \mathrm{sim}(u,v) \cdot \left(r(v,i) - \bar{r}(v) \right)$$

where $\gamma(v,u,i)$ is a predictor of the performance of neighbor v.

Predicting good neighbors

- The predictor γ can be sensitive to the specific target user u, the item i, or any other input of the system.
  - In this work: $\gamma(v,u,i) = \gamma(v)$, i.e., only the neighbor
- We define two predictors, inspired by the clarity score:
  - Item-based user clarity (IUC)
  - User-based user clarity (UUC)

$$\gamma = \gamma(v) = \mathrm{IUC}(v) = \sum_{i \in I} p(i \mid v) \log_2 \frac{p(i \mid v)}{p_c(i)}$$

$$p(i \mid v) = \lambda \frac{rat(v,i)}{5} + (1-\lambda)\, p_c(i)$$

$$p_c(i) = \frac{1}{|I|}$$

$$\gamma = \gamma(v) = \mathrm{UUC}(v) = \sum_{w \in U} p(w \mid v) \log_2 \frac{p(w \mid v)}{p_c(w)}$$

$$p(w \mid v) = \sum_{i : rat(v,i) \neq 0} p(w \mid i)\, p(i \mid v)$$

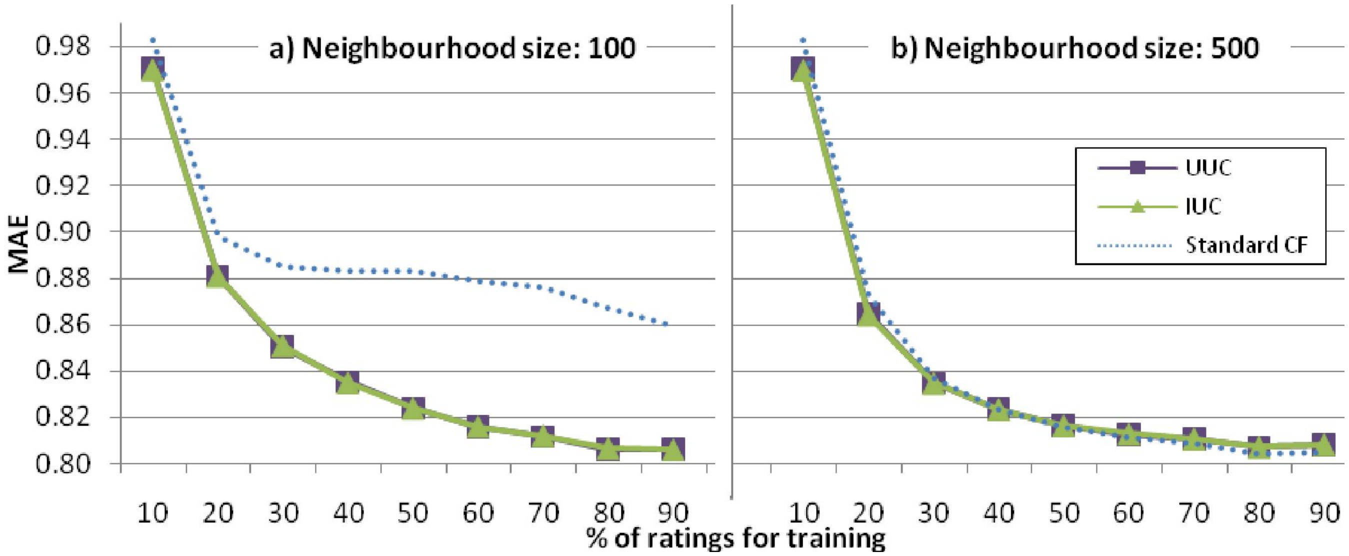
$$p(w \mid i) = \lambda \frac{rat(w,i)}{5} + (1-\lambda)\, p_c(w)$$

$$p_c(v) = \frac{1}{|U|}$$

Evaluation

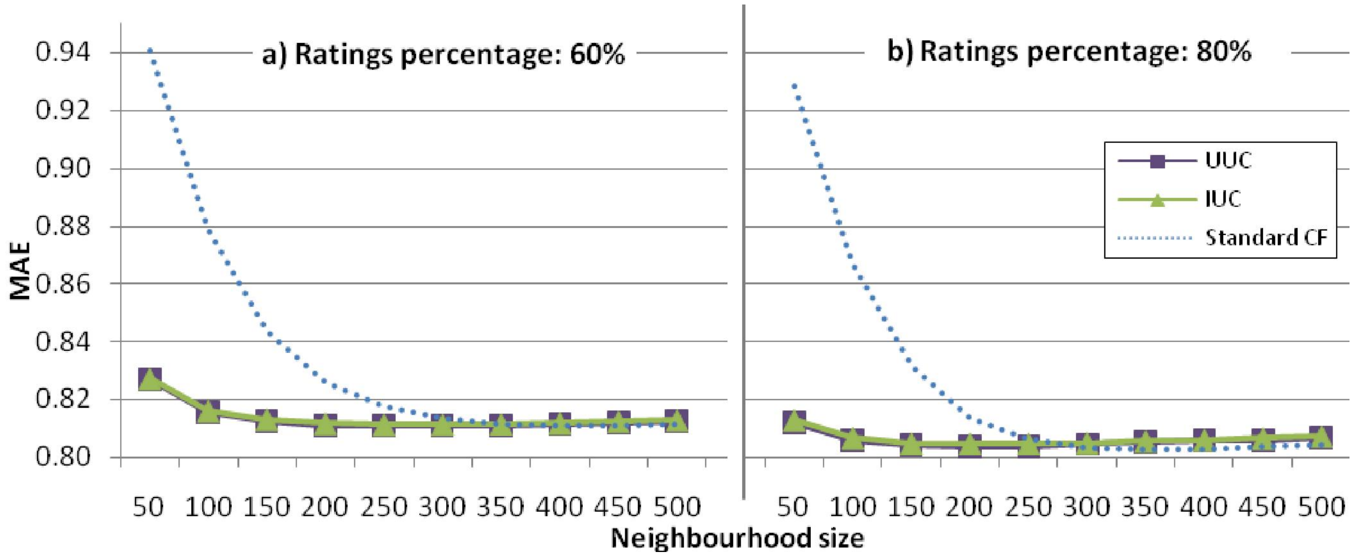
- MovieLens dataset (100k)
- Two variables:
  - Neighborhood size
  - Sparsity (number of available ratings)
- Measure final performance improvements (in terms of MAE) when dynamic weights are introduced

Results I



Performance comparison for different rating density

Results II



Performance comparison for different neighbourhood sizes

Conclusions

- Performance improvements in MAE using dynamic weights for neighbors
- Higher difference for small neighborhoods

- In the future:
  - Other variants of the clarity-based predictor (He & Ounis 2004, Zhou & Croft 2007)
  - Correlation analysis (as in IR)
    - It involves defining user-level performance metrics
  - Extension to other areas: hybrid recommender systems (Adomavicious & Tuzhilin 2005), personalized IR (Castells et al. 2005), rank fusion (Fox & Shaw 1993)

Thank you

Bibliography

- Adomavicius, G., Tuzhilin, A.: Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. IEEE Transactions on Knowledge and Data Engineering 17(6), 734--749 (2005)
- Castells, P., Fernández, M., Vallet, D., Mylonas, P. & Avrithis, Y. (2005), Self-tuning personalized information retrieval in an ontology-based framework, *in* R. Meersman, Z. Tari & P. Herrero, eds, 'SWWS'05: In proceedings of the 1st International Workshop on Web Semantics', Vol. 3762, Springer, pp. 977–986.
- Cronen-Townsend, S., Zhou, Y. & Croft, B. W. (2002), Predicting query performance, *in* 'SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval', ACM Press, New York, NY, USA, pp. 299–306.
- Fox, E. A. & Shaw, J. A. (1993), Combination of multiple searches, *in* 'TREC', pp. 243–252.
- Hauff, C., Hiemstra, D., de Jong, F.: A survey of pre-retrieval query performance predictors. In: 17th ACM conference on Information and knowledge management (CIKM 2008), pp. 1419—1420. ACM Press, New York (2008)
- He, B., Ounis, I.: Inferring query performance using pre-retrieval predictors. In: Apostolico, A., Melucci, M. (eds.) String Processing and Information Retrieval, LNCS, vol. 2346, pp. 43--54. Springer, Heidelberg (2004)
- Zhou, Y., Croft, B.W.: Query performance prediction in web search environments. In: 30th annual international ACM SIGIR conference on Research and development in information retrieval (SIGIR 2007), pp. 543--550, ACM Press, New York (2007)