Does Wikipedia information help Netflix predictions?

John Lees-Miller, Fraser Anderson, Bret Hoehn, Russell Greiner
University of Alberta
{leesmill,frasera,hoehn,greiner}@cs.ualberta.ca

The Netflix Prize

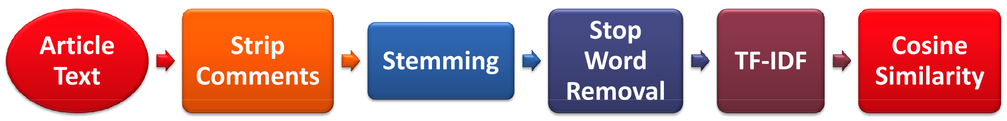
- $1,000,000 to improve their recommender system by 10% based on Root Mean Squared Error (RMSE)
- 100,000,000 ratings from 480,000 users for 17,770 movies

Wikipedia

- Similar movies have similar articles
- Found articles for 93% of the Netflix titles using:
  - Yahoo! web service
  - Edit distance
  - Longest Common Subsequence and Keyword Weighting

Article Similarity

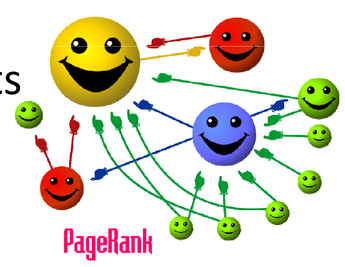
Text similarity:



- Each article becomes a vector of keywords
- TF-IDF measures importance of a keyword in an article
- Cosine similarity: $\cos(a,b) = \frac{a \cdot b}{\sqrt{a \cdot a} \cdot \sqrt{b \cdot b}}$

Hyperlink similarity:

- Article similarity based on common links
- Modify TF-IDF with PageRank: R(q)
- Important links have higher weights



Weighted hyperlink similarity:

- Links are more important if they appear on articles for similar movies
- Movie similarity computed using k-NN

Experiments

- The computed article similarities are used to predict ratings on 1.4m queries using:
- Pseudo-SVD
  - Stochastic gradient descent matrix factorization
  - Proven effective for Netflix recommendations
- k-Nearest Neighbours
  - Linear combination of the ratings given to similar movies
- Blending
  - Combine predictions with others (Reel Ingenuity)

Results

Method	RMSE
k = 24 NN, text similarity	0.97113
k = 24 NN, link similarity	0.96932
k = 24 NN, weighted link similarity	0.96452
PSVD with ratings only (no Wikipedia data)	0.90853
PSVD with link similarity	0.90875
PSVD with weighted link similarity	0.90903
PSVD blends without Wikipedia data	0.88061
PSVD blends with all Wikipedia data	0.88033

- Only 0.0028 improvement with Wikipedia data
- Combining Wikipedia data with ratings data increased prediction error

Conclusions

- Wikipedia articles weakly related to movie similarity
- There are many factors that contribute to similarity: (genre, actors, plot) and it is difficult to extract this from an unstructured source
- Wikipedia information did help some predictions, but by a very small amount

References

Netflix Prize: www.netflixprize.com
PageRank: L. Page, S. Brin, R. Motwani, and T. Winograd, "The PageRank citation ranking: Bringing order to the web," *Stanford Digital Library Technologies Project, Tech. Rep.*, 1998.
Pseudo-SVD: A. Paterek, "Improving regularized singular value decomposition for collaborative filtering," in *Proc. KDD Cup Workshop at SIGKDD'07, 13th ACM Int. Conf. on Knowledge Discovery and Data Mining*, 2007, *pp. 39-42.*

University of Alberta — Department of Computing Science
Alberta Ingenuity Centre for Machine Learning