Advances in Consumer Research
https://acr-journal.com/
Volume-3 | Issue-5 |May 2026
Original Researcher Article

From Credits to Choices: How Creative Team Signals Shape Algorithmic Movie Recommendations and Consumer Discovery.

Mahesh Naik [1]
[1]Department of Basic Sciences and Humanities, SVKM'S NMIMS Mukesh Patel School of Technology Management and Engineering, Mumbai, India
Email: Mahesh.naik@nmims.edu

ABSTRACT
The challenge of connecting consumers with entertainment content that aligns with their tastes represents one of the defining problems of the digital streaming era. This paper presents a multi-method recommendation framework that uses cast and crew metadata as the primary signal for measuring inter-film similarity. Four computational approaches are developed and contrasted: TF-IDF vectorization with cosine similarity, k-Nearest Neighbours (k-NN), a hybrid Decision Tree K-Means classifier, and standalone K-Means clustering. Drawing on a large-scale film credits dataset, the study demonstrates how distinct algorithmic paradigms can extract complementary dimensions of similarity from identical underlying data. Empirical analysis reveals meaningful trade-offs in precision, cluster coherence, and scalability across methods, ultimately suggesting that ensemble deployment outperforms any single approach. The findings carry direct implications for platform designers seeking to improve content discovery and user retention in on-demand entertainment environments...
Keywords: content-based filtering, collaborative metadata, TF-IDF, cosine similarity, k-nearest neighbour, k-means clustering, decision tree, recommendation systems, consumer experience.

INTRODUCTION:

Digital streaming platforms have fundamentally restructured how consumers discover and engage with filmed entertainment. With catalogues numbering in the tens of thousands of titles, the gap between what users want to watch and what they can feasibly browse has widened considerably, making algorithmic recommendation a strategic priority for platform operators and a meaningful factor in consumer satisfaction (Adomavicius & Tuzhilin, 2005).

Existing recommendation architectures can be broadly grouped into two paradigms. Collaborative filtering infers preference from the aggregated behaviour of users with similar taste profiles (Resnick et al., 1994; Breese et al., 1998), whereas content-based filtering derives recommendations from item-level attributes genre, narrative description, or, as in the present study, the identities of the creative professionals involved in production (Lops et al., 2011). Both paradigms carry well-documented limitations: collaborative filtering suffers from cold-start problems and popularity bias, while content-based approaches are constrained by the richness and completeness of available item features.

This paper focuses on a relatively underexplored content signal the composition of a film's cast and directorial team and investigates whether metadata describing creative labour can form the basis of a practically useful recommendation system. Four distinct statistical methods are constructed around a shared feature representation derived from credits data, enabling a direct comparison of their behavioural characteristics. The contribution is threefold: (i) a replicable pre-processing pipeline that transforms raw, semi-structured credits data into machine-readable feature vectors; (ii) a systematic comparative evaluation of four recommendation algorithms on the same corpus; and (iii) empirical insights into the conditions under which each approach succeeds or breaks down.

2. Related Literature

Foundational work on recommender systems has established the theoretical basis for both collaborative and content-based approaches (Adomavicius & Tuzhilin, 2005; Aggarwal, 2016). The application of vector-space models—particularly TF-IDF representations—to information retrieval tasks is comprehensively treated in Manning, Raghavan, and Schutze (2008), and their adoption within recommendation contexts is well established in the content-based filtering literature (Lops et al., 2011).

The use of k-NN as a proximity-based recommender is foundational to neighbourhood-based collaborative filtering (Breese et al., 1998), though its application to content vectors in high-dimensional spaces raises documented concerns about distance concentration. K-Means clustering has been used to organize item spaces in recommendation settings (Arthur & Vassilvitskii, 2007), while hybrid architectures that chain unsupervised clustering with supervised classification represent a more recent development aimed at improving computational efficiency without sacrificing personalization quality. Matrix factorization techniques (Koren et al., 2009) represent an alternative direction not pursued here, as our

focus is on interpretable, metadata-driven methods rather than latent-factor models.

3. Data and Pre-processing

3.1 Dataset

The empirical foundation of this study is a publicly available film credits dataset ("credits.csv"), sourced from The Movies Dataset hosted on Kaggle, which itself draws on metadata aggregated from The Movie Database (TMDb). The dataset spans theatrical releases across multiple decades and production geographies, providing broad coverage of mainstream Hollywood output alongside a smaller representation of international and independent cinema.

In its raw form, the dataset contains 45,439 records distributed across six fields. Two fields carry the primary analytical content: cast and crew. Each record in these fields is stored as a serialized string representation of a Python list of dictionaries, where every dictionary encodes a single credited individual along with associated attributes. For cast members, these attributes include a unique cast identifier, the character name portrayed, the billing order (cast_id), and the performer's full name. For crew members, attributes include a unique credit identifier, the individual's department (e.g., Directing, Production, Camera), their specific job title (e.g., Director, Producer, Director of Photography), and their full name. Two further fields provide a numeric film identifier (id) and a supplementary identifier linking records to other tables in the broader dataset.

The remaining two derived fields actors and directors are not present in the raw file but are engineered during pre-processing (detailed in Section 3.2).

Table 1. Structure of the credits.csv dataset

Field	Type	Description
`cast`	String (serialized list)	Full cast roster with character names, billing order, and performer identities
`crew`	String (serialized list)	Full crew roster with department, job title, and individual identities
`id`	Integer	Unique numeric film identifier
`actors`	Derived text	Space-separated names of up to five top-billed cast members (engineered)
`directors`	Derived text	Space-separated name(s) of credited director(s) (engineered)
`metadata`	Derived text	Concatenation of `actors` and `directors` fields; primary modelling input (engineered)

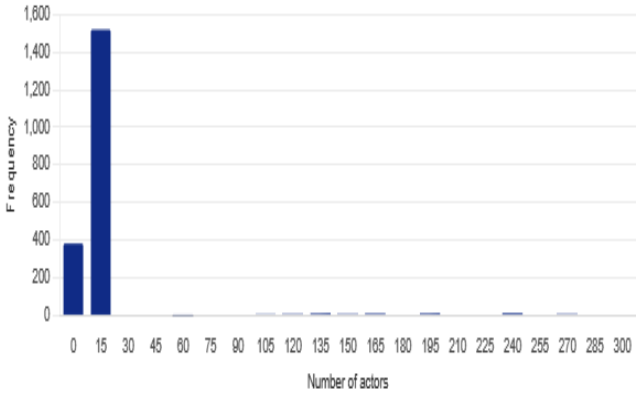

Figure 1. Distribution of actors per film.

The histogram reveals a pronounced right skew, with the majority of productions featuring fewer than 30 credited performers and a small number of large ensemble productions forming a long upper tail. This distribution motivates the decision to cap actor extraction at five names during pre-processing.

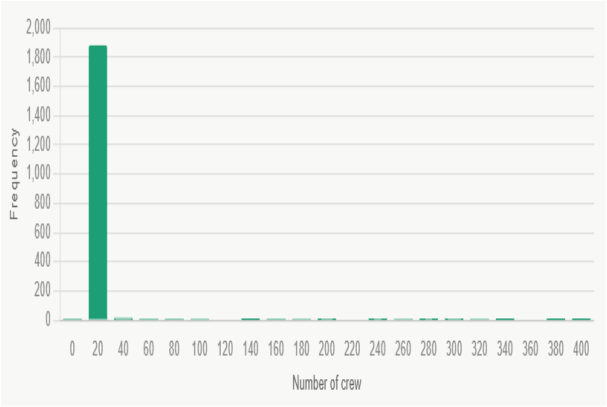

Figure 2. **Distribution of crew members per film.**

The dataset skews toward English-language productions, particularly films distributed through major North American studios from the mid-twentieth century onward. Silent-era and early sound films are represented through the prominence of figures such as John Ford and Georges Méliès in the director frequency rankings, indicating that historical depth is present, though uneven. International productions from non-Anglophone industries including Indian, French, and Japanese cinema appear in smaller proportions, and credits documentation for these entries is less consistently complete, which may affect the quality of their vector representations in the TF-IDF space.

Data quality. Of the 45,439 raw records, zero missing values are observed across the cast, crew, and id fields following initial loading, suggesting that the upstream data curation process applied completeness filters prior to release. Duplicate records are nonetheless present and removed during pre-processing. The principal data quality challenge lies not in missingness but in the variable

granularity of credits: some productions list dozens of named cast members with full character attribution, while others particularly older or lower-budget films carry minimal cast information. This heterogeneity in record depth is a structural feature of industry crediting practices rather than an artifact of dataset construction, and it motivates the decision to cap actor extraction at five names and to rely on TF-IDF weighting to discount the influence of individuals who appear across an unusually large number of productions.

3.2 Cleaning and Parsing

Prior to any feature engineering, records with absent cast or crew entries were removed using listwise deletion, and exact duplicate rows were eliminated. The serialized string fields were then parsed into native Python data structures using ast.literal_eval(), restoring the relational structure necessary for extracting individual names.

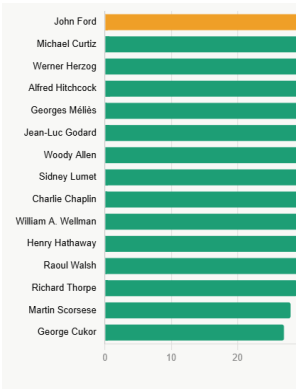

Figure 3. Top 15 most frequently appearing directors

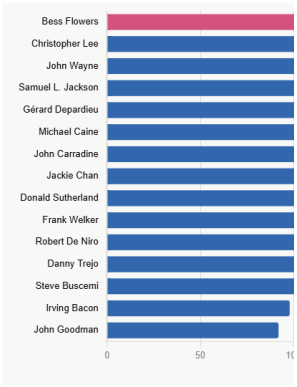

Figure 4. Top 15 most frequently appearing actors

3.3 Feature Extraction

Two extraction functions operationalize the creative team concept. First, for each film, the names of up to five billed cast members are retrieved in billing order—a design choice justified by the right-skewed distribution of cast sizes observed during exploratory analysis, where a small number of blockbuster productions inflate the upper tail. Second, the director's name is extracted from crew records by filtering on the job title field. Actor and director names are then concatenated into a single free-text "metadata" field that functions as the primary input to all downstream models.

Two additional numerical features are derived for use in the Decision Tree model: the total count of credited cast members (num_actors) and the total count of crew members (num_crew).

3.4 Sampling

To balance analytical tractability with representativeness, a random subsample of 2,000 films is drawn from the full dataset using a fixed random seed, enabling reproducibility across experimental runs.

4. Exploratory Data Analysis

Before modelling, distributional properties of the dataset were examined to motivate subsequent feature engineering choices.

Cast size distribution. The histogram of actors per film is strongly right-skewed, with the majority of productions featuring relatively compact casts and a small number of ensemble or ensemble-adjacent productions creating a long upper tail. This distributional shape justifies capping the actor extraction at five names: including the full cast would introduce noise from peripheral performers while contributing little signal about a film's creative identity.

Crew size distribution. Crew size exhibits an even more pronounced positive skew than cast size, reflecting the wide range of production scales represented in the dataset. Large-budget studio productions with extensive technical departments appear as outliers relative to the central mass of smaller productions.

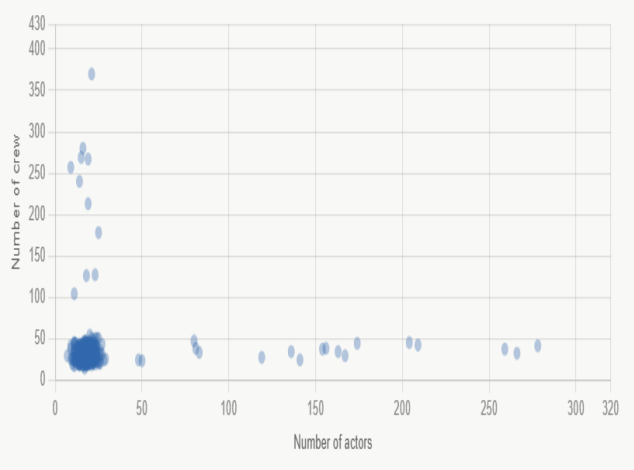

Figure 5. Cast size vs. crew size (production scale)

Actor and director frequency. Analysis of the most frequently appearing names confirms that a small cohort of prolific performers and directors—including figures such as Bess Flowers, Christopher Lee, and John Ford—account for a disproportionate share of appearances. This concentration creates a potential source of bias in similarity metrics, as co-presence of a ubiquitous actor may generate spurious similarity between otherwise unrelated films. The TF-IDF weighting scheme applied in modelling is specifically designed to mitigate this effect by down weighting terms that appear frequently across the corpus.

Actor-crew size correlation. A scatter plot of the two numerical features reveals a positive, though imperfect, association between cast size and crew size, consistent with both being proxies for production scale. The correlation is strong enough to suggest that one variable

partially subsumes the other, a consideration that bears on the discriminatory power of the Decision Tree model.

5. Methodology

5.1 Shared Feature Representation: TF-IDF Vectorization

All four models operate on a shared feature substrate derived by applying TF-IDF vectorization to the metadata field. TF-IDF encodes each film as a sparse vector in which the weight assigned to each creative contributor's name reflects how distinctive that name is to that particular film relative to the full corpus (Manning et al., 2008). Common names—those appearing across many films—receive lower weights, while names that are concentrated in a small number of productions receive higher weights. Stop words are removed during tokenization. The result is a high-dimensional sparse matrix in which rows correspond to films and columns to unique contributor names.

5.2 Model 1: TF-IDF with Cosine Similarity

The simplest recommendation pathway computes pairwise cosine similarity across all film vectors in the TF-IDF matrix. Cosine similarity measures the angle between two vectors rather than their Euclidean distance, making it insensitive to differences in vector magnitude—a desirable property when comparing films that vary substantially in the number of credited individuals. For a target film, the system ranks all other films by descending similarity score and returns the top-N as recommendations.

5.3 Model 2: k-Nearest Neighbours

The k-NN model uses the same TF-IDF representation but identifies similar films through a fitted neighbourhood structure rather than full pairwise comparison. With k set to 6 (inclusive of the query film), the model retrieves the five most proximate neighbors in feature space using cosine distance as the proximity metric. The k-NN approach emphasizes local structure—the configuration of films in the immediate vicinity of the query—and may surface recommendations that a global similarity ranking would rank lower.

5.4 Model 3: Decision Tree Classifier with K-Means Cluster Labels

The third model combines unsupervised and supervised learning in a two-stage pipeline. In the first stage, K-Means clustering (k=5) is applied to the TF-IDF matrix to assign each film to one of five clusters. These cluster assignments are then used as class labels in the second stage, where a Decision Tree Classifier is trained to predict cluster membership from the two numerical features: num_actors and num_crew. At inference time, the classifier predicts the cluster of the query film from its numerical features, and recommendations are drawn from other films assigned to the same predicted cluster. This architecture is motivated by computational efficiency: generating recommendations requires only a simple decision tree lookup rather than proximity search in a high-dimensional space.

5.5 Model 4: K-Means Clustering

The fourth model applies K-Means directly to the TF-IDF matrix with five cluster centres. Each film is assigned to its nearest centroid, and recommendations for a query film are drawn from other members of its cluster. K-Means makes no use of individual pairwise similarity at inference time, instead treating cluster co-membership as a sufficient condition for recommendation. This approach is scalable to large corpora but, as discussed below, is sensitive to the geometry of the feature space.

6. Results and Inference

6.1 Cosine Similarity Matrix

Visualization of the full pairwise cosine similarity matrix as a heat map reveals that the overwhelming majority of film pairs register near-zero similarity scores, producing a predominantly uniform low-intensity field punctuated by a bright diagonal representing each film's self-similarity of 1.0. Sparse off-diagonal bright points correspond to small subsets of films sharing significant cast or crew overlap—likely franchise entries, director retrospectives, or productions drawing repeatedly from the same ensemble. This extreme sparsity validates the use of similarity-based retrieval: meaningful connections exist in the data, but they are rare and concentrated, making targeted search preferable to broad-brush approaches.

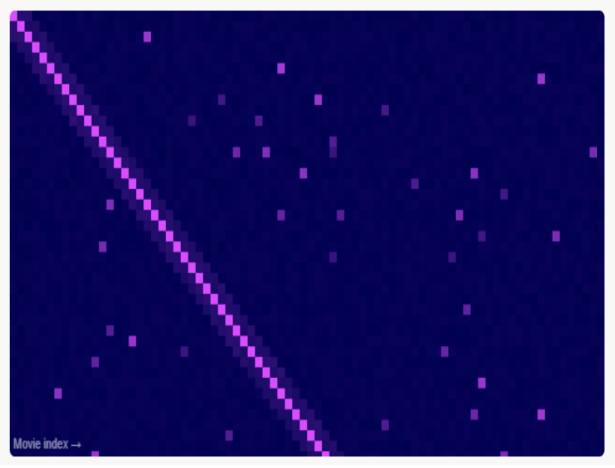

Figure 6. Cosine similarity matrix (TF-IDF vectors)

Each cell shows pairwise cosine similarity between two films. Bright diagonal = self-similarity (1.0); sparse off-diagonal spots indicate rare meaningful overlaps.

6.2 K-Means Cluster Distribution

The distribution of films across the five K-Means clusters is severely imbalanced. Cluster 0 absorbs approximately 1,350 of the 2,000 sampled films, while the remaining four clusters range from roughly 75 to 230 members each. This outcome reveals a structural limitation of K-Means in high-dimensional sparse spaces: when the majority of films are equidistant from one another in TF-IDF space, the algorithm defaults to agglomerating them into a single large residual cluster rather than partitioning the space meaningfully. The smaller clusters, by contrast, exhibit

greater internal coherence and likely represent genre-specific or production-company-specific groupings with identifiable creative team signatures. The practical implication is that recommendations generated from Cluster 0 will be weakly differentiated, while those from the smaller clusters may be quite precise.

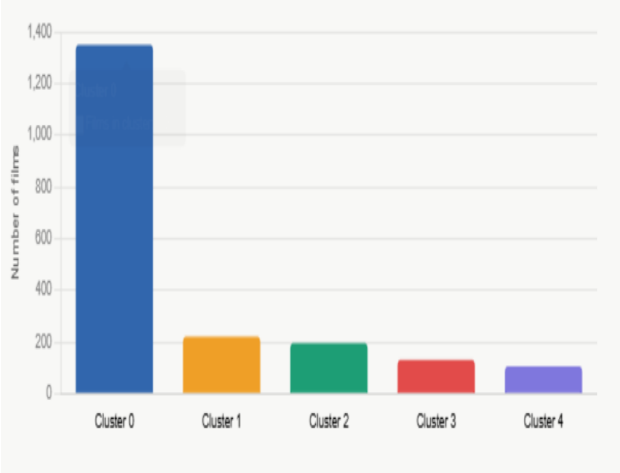

Figure 7. K-Means cluster distribution (k = 5)

6.3 k-NN Distance Profile

Examination of the cosine distances between a sample query film and its five nearest neighbors reveals that all neighbors lie at distances above 0.8, on a scale where 0 denotes identity and 1 denotes complete orthogonality. The distances increase gradually from the first to the fifth neighbor (approximately 0.80 to 0.87), confirming that the feature space is sparse but not structureless: the k-NN algorithm can establish a meaningful gradient of proximity even when absolute similarity values are low. The high distances do, however, signal that even the closest recommended films share relatively little cast-and-crew overlap with the query title, which has implications for perceived recommendation quality from a consumer perspective.

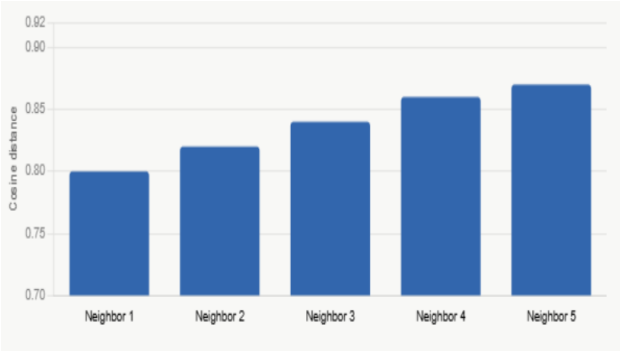

Figure 8. k-NN cosine distances to five nearest neighbours

6.4 Comparative Model Behaviour

The TF-IDF cosine model and the k-NN model produce identical top-5 recommendations for the test query, reflecting their shared feature representation and metric. The Decision Tree and K-Means models produce divergent recommendations, attributable to the coarser granularity of cluster-based retrieval relative to continuous-similarity ranking. The Decision Tree's reliance on only two numerical predictors—cast size and crew size—to proxy the rich textual signal captured by K-Means clustering introduces further approximation error, likely explaining the systematic divergence between tree-based and direct K-Means recommendations observed in the output.

7. Discussion

7.1 Theoretical Implications

The results collectively illustrate that creative team composition is a viable, if imperfect, basis for content-based recommendation. The sparsity of the similarity matrix suggests that cast-and-crew overlap is a relatively rare property in the full film corpus, but when it does occur it tends to cluster in ways that are theoretically interpretable—franchise series, auteur filmographies, repertory ensembles. This is consistent with consumer research on the role of parasocial relationships with stars and directors in shaping viewing preferences (Aggarwal, 2016).

The cluster imbalance observed with K-Means raises broader questions about the applicability of centroid-based methods to high-dimensional text data. Density-sensitive clustering approaches, such as DBSCAN or hierarchical agglomerative methods, may better accommodate the irregular geometry of TF-IDF spaces and warrant investigation in future work.

The hybrid Decision Tree K-Means model highlights a fundamental tension in recommendation system design: the desire for computational efficiency pushes toward simpler feature representations, but recommendation quality depends on the fidelity with which those representations capture genuine item similarity. Two numerical proxies for production scale are insufficient to replicate the discriminative power of full TF-IDF vectors.

7.2 Consumer Experience Implications

From a consumer behaviour standpoint, the choice of recommendation method is not merely a technical decision but a design choice with measurable implications for user experience. Models that produce highly similar recommendations (TF-IDF cosine and k-NN) may satisfy users seeking familiar creative territory—fans of a particular actor or director—while cluster-based methods may serve users open to broader exploration within a loosely defined taste neighbourhood. Platforms might therefore consider surfacing recommendations from multiple algorithmic sources simultaneously, presenting them in distinct interface zones to serve different discovery intents.

7.3 Limitations

Several limitations bound the generalizability of these findings. First, the system uses cast and crew metadata exclusively, omitting narrative content, genre classifications, critical reception, and user behavioural data—all of which likely carry substantial predictive signal for consumer preference. Second, the equal weighting of the top-five billed actors does not account for

the variable centrality of each performer to a film's identity; a brief cameo appearance by a prolific star can inflate similarity scores inappropriately. Third, the dataset underrepresents international and independent cinema, where credits documentation is less complete, potentially disadvantaging these films in similarity search.

8. Conclusion

This paper has developed and compared four statistical approaches to movie recommendation grounded in cast-and-crew metadata. The findings demonstrate that no single method dominates across all evaluation criteria: TF-IDF cosine similarity and k-NN offer the most interpretable and consistent pairwise recommendations, K-Means provides a computationally efficient cluster structure but suffers from severe imbalance, and the Decision Tree hybrid achieves efficiency at the cost of fidelity to the underlying text-based similarity signal. Taken together, these results argue for multi-method deployment strategies in which different algorithmic perspectives are combined to mitigate the weaknesses of any individual approach.

Future research should incorporate additional feature modalities plot embedding's, genre taxonomies, audience ratings—and explore evaluation frameworks that measure recommendation quality through consumer-facing metrics such as click-through rates, completion rates, and stated satisfaction. Addressing the cold-start problem for new releases and emerging talent represents a particularly important practical challenge for continued work in this domain.

.

REFERENCES

1. Adomavicius, G., & Tuzhilin, A. (2005). 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.

2. Aggarwal, C. C. (2016). Recommender systems: The textbook. Springer.

3. Arthur, D., & Vassilvitskii, S. (2007). k-means++: The advantages of careful seeding. Proceedings of the 18th Annual ACM-SIAM Symposium on Discrete Algorithms, 1027–1035.

4. Bell, R. M., Koren, Y., & Volinsky, C. (2007). The BellKor solution to the Netflix Prize. Netflix Prize Documentation.

5. Breese, J. S., Heckerman, D., & Kadie, C. (1998). Empirical analysis of predictive algorithms for collaborative filtering. Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, 43–52.

6. Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8), 30–37.

7. Lops, P., De Gemmis, M., & Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In F. Ricci, L. Rokach, B. Shapira, & P. B. Kantor (Eds.), Recommender systems handbook (pp. 73–105). Springer.

8. Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

9.

10. Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. Advances in Neural Information Processing Systems, 26.

11. Paterek, A. (2007). Improving regularized singular value decomposition for collaborative filtering. Proceedings of KDD Cup and Workshop.

12. Resnick, P., Iacovou, N., Suchak, M., Bergstrom, P., & Riedl, J. (1994). GroupLens: An open architecture for collaborative filtering of netnews. Proceedings of the 1994 ACM Conference on Computer Supported Cooperative Work, 175–186.

..