Applying Collaborative Filtering Techniques to Movie Search for Better Ranking and Browsing

Seung-Taek Park and
David M. Pennock
(ACM SIGKDD 2007)

1

Introduction

- Two types of technologies are widely used to overcome information overload: information retrieval and recommender systems.
- Information retrieval systems accept a query from a user and return the user relevant items against the query.
- Information retrieval systems work somewhat passively while recommender systems look for the need of a user more actively.

2

Introduction

- Recommender systems predict the need of a user based on his historical activities and recommend items that he may like even though the user does not specifically request it.

3

Introduction

- We propose a new approach to combine informational retrieval and recommender system for better search and browsing.
- We use collaborative filtering algorithms to calculate personalized item authorities in search.

4

Introduction

- To demonstrate our approach, we build a prototype personalized movie search engine called MAD6.
- MAD6 combines both information retrieval and collaborative filtering techniques for better search and navigation.

5

Introduction

- The ranking of returned items in web search engines is the combination of the item proximity and authority.
- Item proximity, sometimes called item relevance, denotes the item's similarity or relevance to the given query.
- Item authority denotes the importance of a item in the given item set.

6

Ranking Algorithm

- Navigational queries : when users already know what they are looking for.
- Informational queries : when a user searches for something, in many cases he does not know much about the object.

7

Item Proximity

- We build our own database of more extensive metadata for better recall in search results.
- In addition to movie titles, we index a large amount of metadata including genres, names of actors, directors, characters, plots, MPGA ratings, award information, reviews of critics and users, captions from trailers and clips, and so on.

8

Item Proximity

- We conducted a test comparing our extensive indexing system with IMDB and Yahoo! Movies current search.
- We downloaded movie data from IMDB and generated queries for the 100 most popular movies, where popularity is measured by the number of user ratings.

9

Item Proximity

- We use five movie metadata fields: names of actors, directors, characters, plots, and genres.
- The two highest TF/TFIDF words of each of the top 100 popular movies are selected as a query and only the top 10 returned movies are analyzed.

10

Item Proximity

Table 1: Hit ratios of three movie search engines for the top 100 most popular movies; Only the top 10 returned movies are considered. "DB" denotes our base system with an extensive index. Two top TF/TFIDF terms from five metadata, including names of actors, directors, characters, plots, and genres, are selected as a query for each movie in the top 100 popular movies. Then each query is submitted to three systems, IMDB, Yahoo! Movies and our base system. The popularities of movies are measured by the number of user ratings. We downloaded the IMDB movie content data and conducted this test in April 2006.

	TF		TFIDF	
	HIT	No Returns	HIT	No Returns
IMDB	4	2	6	2
current Yahoo!	2	94	2	95
DB	**33**	25	**37**	43

11

Item Proximity

- We find that users often provide some extra information of items which do not exist in our database.
- "jlo" - the nick name of actress Jennifer Lopez - is often found in the users' reviews of the movies she has starred.

12

Item Proximity

- We use the Yahoo! Search API for getting web information.
- Each time our system gets a query from a user, it conducts a site-limited search through the API and extracts information of corresponding items from our database.

Item Proximity

$$Web(i,q) = \frac{(N+1-L(i,q))}{N} * \gamma$$

- L(i , q) is the highest rank of an item i in the web search result for the query q.
- N and γ are the maximum number of returns from the search engine and a normalized factor.
- We set γ= 13 and N = 50.

Item Proximity

Table 2: The effect of web relevance. "jlo" is submitted to each system as a query. Bold represents items relevant to Jennifer Lopez. The results are as of April 2006.

System	Top 10 movie results
Yahoo!	No returns
IMDB	1. Fejfys (1968) 2. Mihajlo Bata Paskaljevic (2001) 3. Mihajlo Petrovic Alas (1968) 4. Mijlocas la deschidere (1979) 5. Scopul si mijloacele (1983) 6. A Szr s gykr fejldse (1961) 7. Ziveo zivot Tola Manojlovic (1973) 8. Kumonosu j (1957) 9. The Hi-Lo Country (1998) 10. Vlemma ton Odyssea, To (1995)
DB	No returns
Web	1. **Maid in Manhattan (2002)** A+ 2. **Angel Eyes (2001)** A- 3. Let's Dance (1950) B+ 4. Sweet 15 (1996) B+ 5. My Family (1995) B+ 6. **U-Turn (1997)** B+ 7. **The Cell (2000)** B 8. **The Wedding Planner (2001)** C-

Item Proximity

$$Prox(i,q) = max(Web(i,q), DB(i,q))$$

- DB(i , q) and Web(i , q) denote DB and Web relevancies of an item i for the given query q.

Item Proximity

$$Auth_i = \frac{\overline{r_i} + log_\gamma |U_i| + \overline{c_i} + log_{10}(10 * aw_i + 5 * an_i)}{\delta}$$

U_i is the set of users who have rated item i, $\overline{r_i}$ is the average rating of item i over all users, $\overline{c_i}$ is the average critic rating of item i, aw_i is the number of awards that item i has won, an_i is the number of awards that item i has been nominated for, and δ is a normalization factor such that the maximum global item authority is 13. we set γ such that the maximum value of $log_\gamma |U_i|$ is 13.

Item Proximity

$$sim(i,j) = \frac{\sum_{u \in U}(r_{u,i} - \overline{r}_u) \cdot (r_{u,j} - \overline{r}_u)}{\sqrt{\sum_{u \in U}(r_{u,i} - \overline{r}_u)^2} \cdot \sqrt{\sum_{j \in U}(r_{u,j} - \overline{r}_u)^2}}$$

$$sim'(i,j) = \frac{min(|U_i \cap U_j|, \gamma)}{\gamma} * sim(i,j)$$

- $r_{u,i}$ is the rating of user u for item i and r_u is user u's average item rating.
- U_i denotes a set of users who have rated the item i. We set γ = 50.

Prediction

$$p_{u,i} = \overline{r}_i + \frac{\sum_{j \in I_u} sim'(i,j) * (r_{u,j} - \overline{r}_j)}{\sum_{j \in I_u} |sim'(i,j)|}$$

- We use user rating information from Yahoo! Movies to calculate item similarities.
- r_i and I_u denote the average rating of the item i over all users and a set of items the user u has rated.

19

Ranking Score Function

$$MADRank(i,q,u) = \alpha * Auth(i,q,u) + (1-\alpha) * Prox(i,q)$$

- α is a weighting factor for item authorities.
- We set α = 0.5.
- We set the MADRank score to 13 if the title of an item exactly matches to the given query.

20

Ranking System

Table 4: Top 10 results of different ranking methods for the query "arnold action". The results are as of April 2006.

Ranking	Top 10 movie results
current Yahoo!	No items return
IMDB	1. Einleitung zu Arnold Schoenbergs ... (1973) 2. Benedict Arnold: A Question of Honor (2003) 3. Love and Action in Chicago (1999) (V) 4. Mary-Kate and Ashley in Action! (2001) 5. Armed for Action (1992) 6. Demonstrating the Action of the ... (1900) 7. Peace Is Every Step: Meditation ... (1998) 8. Rock 'n' Roll Space Patrol Action Is Go! (2005) 9. Leitstandarte SS-Adolf Hitler im .. (1941) 10. Action Figures: Real and Unreal (2005) (V)
DB	1. Arnold Schwarzenegger DVD 2-Pack - The Sixth Day/The Last Action Hero(2003) 2. THE LAST ACTION HERO (1993) and the 2 - DVD Special Edition of THE 6TH DAY 3. Warner Home Video DVD Action 4-Pack (1997) 4. Last Action Hero (1993) 5. The 6th Day (2000) 6. Eraser (1996) 7. Commando (1985) 8. True Lies (1994) 9. Nancy Drew - A Haunting We Will Go (1977) 10. Out for Justice (1991)

Web	1. Arnold Schwarzenegger DVD 2-Pack - The Sixth Day/The Last Action Hero(2003) 2. Last Action Hero (1993) 3. Commando (1985) 4. End of Days (1999) 5. Eraser (1996) 6. True Lies (1994) 7. Terminator 2 - Judgment Day (1991) 8. Raw Deal (1986) 9. Terminator 3: Rise of the Machines (2003) 10. Collateral Damage (2002)
GRank	1. True Lies (1994) 2. Last Action Hero (1993) 3. Commando (1985) 4. Terminator 2 - Judgment Day (1991) 5. End of Days (1999) 6. Eraser (1996) 7. The Terminator (1984) 8. The Bridge on the River Kwai (1957) 9. Terminator 3: Rise of the Machines (2003) 10. The Fugitive (1993)
PRank	1. Terminator 2 - Judgment Day (1991) 2. Commando (1985) 3. True Lies (1994) 4. Last Action Hero (1993) 5. The Terminator (1984) 6. T2 The Ultimate Edition DVD (1991) 7. The Bridge on the River Kwai (1957) 8. Blockbuster (1988) 9. Total Recall (1990) 10. The Fugitive (1993)

21

Offline Evaluation

- We use search click data from Yahoo! Search containing search queries and clicked URLs.
- If a URL is a Yahoo! Movies page, we extract the yahoo movie id from the URL and find the movie title in our database.
- If a URL is an IMDB page, we submit the URL to Yahoo! Search and find the title, then find the matching Yahoo! movie id.

22

Offline Evaluation

- If we cannot find a corresponding movie by title match, we submit a query to Yahoo! Search and select the first returned movie as its counterpart in Yahoo! Movies.

23

Offline Evaluation

$$HR = \frac{|H|}{|N|} \qquad ARHR = \frac{1}{|N|} \sum_{i \in H} \frac{1}{r_i}$$

- N is a set of test instances and H is a set of hit instances within the top 10 results.
- r_i is the actual rank of the target movie i.

24

Offline Evaluation

Table 6: Offline experiment results.

Metric	IMDB	Yahoo!	DB	Web	Grank
HR	.7435	*.5241*	.5879	.8086	**.8155**
ARHR	.6544	.4615	*.4479*	**.7585**	.7303

- Hit rate is used to capture recall of search results.
- Average reciprocal hit rank is used to measure the quality of search results.
- Our offline test is biased to Web ranking, since we use Yahoo! Search click data.
- Our test may be biased to IMDB, since most URLs in our test data come from IMDB.

25

MAD6

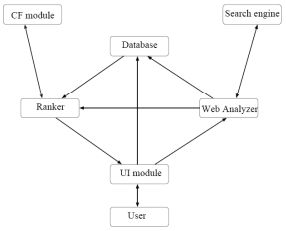

Figure 1: The architecture of MAD6.

26

MAD6

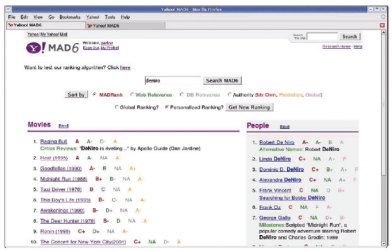

Figure 2: Search Result

27

Online Evaluation

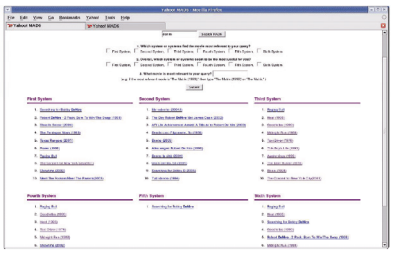

Figure 3: Test demo

28

Online Evaluation

Table 7: Online experiment results. The "Recall" column lists the percentages of users who selected the corresponding system in answering question 1. The "Quality" column lists the percentages for question 2.

Test group	Systems	Recall	Quality
All (180 feedbacks from 44 users)	IMDB	*22.2*	15.6
	Yahoo!	22.8	*13.9*
	DB	36.7	20
	Web	52.2	38.3
	GRank	53.9	**40**
	PRank	**57.8**	38.3
Navigational queries (49 feedbacks from 23 users)	IMDB	42.9	34.7
	Yahoo!	53.1	30.6
	DB	*44.9*	*22.4*
	Web	57.1	34.7
	GRank	59.2	28.6
	PRank	**63.3**	**36.7**
Informational queries (131 feedbacks from 40 users)	IMDB	14.5	8.4
	Yahoo!	*11.5*	*7.6*
	DB	33.6	19.1
	Web	50.4	39.7
	GRank	51.9	**44.3**
	PRank	**55.7**	38.9

29

Conclusions

- In this paper, we discuss our new ranking method, which combines recommender systems and search tools for better informational search and browsing.
- In both offline and online tests, MAD6 seems to provide users better search recall and quality than IMDB search and Yahoo! Movies current search by combining proximities and authorities of the returned items.
- Even though MAD6 is one application in the movie domain, we believe that our approach is general enough to apply other domains including music, travel, shopping and web search.

30