Evaluating the use of Semantics in Collaborative Recommender Systems: A User Study

Patricia Kearney[1], Mary Shapcott[1], Sarabjot S. Anand[2], David Patterson[1]
[1]University of Ulster, School of Computing and Mathematics, Shore Road, Newtownabbey, Co. Antrim, BT37 0QB
{kearney-p3, cm.shapcott, wd.patterson}@ulster.ac.uk
[2]Department of Computer Science, University of Warwick, Coventry, England, CV4 7AL, s.s.anand@warwick.ac.uk

Abstract

In this paper we report on a pilot user study aimed at evaluating two aspects of recommender systems that have not been the aim of previous user studies in the field. Firstly, item semantics may be incorporated into a collaborative recommender system and we wish to measure the effect on user satisfaction. Secondly, we would like to evaluate different approaches to collecting ratings from users: the ratings that are used to seed their profile with a collaborative filtering system. Key indications from the study are: users do prefer recommendations generated by semantically enhanced recommender systems; the user's satisfaction with a recommendation set is different from the sum of their satisfaction with the individual items with the set and the approach to collecting item ratings from the user should be tailored to the algorithm being used. Finally, recommender systems within the movie domain seem to be more useful for "movie buffs" rather than the "average movie watcher" for whom recommending simply the most popular movies seems to be most appropriate.

1 Introduction

As the amount of information on the Internet continues to expand users continue to need help in finding what they want. The problems associated with information overload are not new and are often typified by the difficulties encountered in the search for useful information on the web (Anand and Mobasher 2005). The issues have been widely discussed both in the context of information retrieval (Belkin and Croft 1992) and recommender systems research (Borchers et al. 1998).

To address the overload problem recommender systems have been developed for many domains, from books to movies and from email to e-tailers, to aid users in filtering the domain's content to reach the content which will be of most interest to them.

A taxonomy of the different approaches and their various implementations, both past and present, is presented in Adomavicius and Tuzhilin (2005). Common shortcomings (Burke 2002) of the different types of recommender systems include the new user/new item problem, lack of recommendation diversity and sparsity of ratings.

One approach which has provided useful results in addressing some of these shortcomings has been the integration of domain knowledge into the recommendation process. Ziegler et al. (2005), for example, utilizes taxonomic information on products to classify products, by topics, to address rating information sparsity. Cho and Kim (2004) apply a product taxonomy to address scalability issues.

Mobasher, Jin and Zhou (2004) proposed the incorporation of semantic knowledge about items in their item-based collaborative filtering. An item ontology is integrated into the user profile generation process in work by Middleton, Shadbolt and Roure (2004), Anand, Kearney and Shapcott (2007) and Haase et al. (2005), resulting in an improvement in accuracy of recommendations through a deeper understanding of user behaviour with respect to the underlying domain. Furthermore, Anand, Kearney and Shapcott (2007) argue that different user contexts can be identified using semantic information and that this, in turn, can be used to generate more accurate recommendations.

The "new-user" problem has also been tackled using ontologies (Middleton, Shadbolt and Roure 2004) and the use of a seed (predictor) set of ratings from the user at registration as required by MovieLens and Jester (Gupta et al. 1999). While MovieLens allows the user to choose the movies to rate, Jester has a fixed set of jokes that must be rated by all users.

Although there is increasing interest in the use of user studies in evaluating recommender systems (see Section 5), to the best of our knowledge, no user study has been carried out with the aim of evaluating

- recommendations generated using additional semantic information
- whether the context of a user's visit can be discovered through the incorporation of semantic knowledge and whether by knowing the context, the desirability of recommendation sets can be improved[1]
- what is the best approach to getting the new user to provide seed ratings to the recommendation system

[1] In keeping with (Anand, Kearney and Shapcott., 2007), rather than trying to describe context in terms of a set of features associated with the type of device, location and date/time, we model context as a hidden process that at any time can be in one of a finite set of states that have a bearing on the user's behavior

In this paper we report on findings from a pilot user study which has been designed with the aim of evaluating both the usefulness of item semantics and various approaches to obtaining seed ratings in collaborative recommender systems. Additionally, we discuss the results of an observational study with fifteen users in the movie domain.

The rest of the paper is organized as follows. In Section 2 the findings of the observational study are presented. In Section 3 we describe the design and architecture of the proposed user study. In Section 4 we present the results of an initial pilot study which uses profiles derived from real users carrying out specified browsing tasks. Section 5 covers related work in the field. In Section 6 we conclude the paper by summarizing key findings and outlining future work to be carried out to extend this research.

2 Observational study

An observational study allows the observer to view what users actually do in context (Preese et al., 1994). Before embarking on a user study we felt it was important to gather information about user motivation and browsing behaviour. Observing users is an effective aid to designing clear tasks for use in recommender evaluation as systems may end up being used in different ways than expected (Herlocker et al. 2004).

An initial observational study was carried out with fifteen users recruited from the Computer Science department at the University of Ulster. To participate in the observational study, the users had to have prior experience of using some online movie retailer's web site for browsing and/or purchasing movies. The demographics of the user group are shown below in Table 1.

Age Range	20-30	30-40	40-50	50-60
Male	3	4		
Female	3	2	2	1

Table 1 Demographics of observational study

Each user was initially asked to browse an online movie retailer looking for web pages that interested them. As they browsed we observed and videoed them and, where applicable, asked them questions aimed at eliciting the motivations behind their browsing behaviour. From observing users' browsing behaviour on the movie site we gained information as to which attributes most influenced the fifteen users to watch or purchase a movie. These are presented in Table 2.

As shown in Table 2 many users quoted the same influences. Several pieces of semantic information were repeatedly quoted by users as having significant influence on their decision to watch or purchase a movie. These included the genre of the movie, the plot, external reviews/user ratings, actors, directors and the release date of the movie.

Having carried out the initial browsing task, the user was asked to register with the well-known movie recommender site, MovieLens (http://www.movielens.org) and to rate at least the minimum number of movies required (fifteen movies). The users then looked at the recommendations provided ("Top Picks for You"). The purpose of this exercise was twofold – first to ascertain whether they had already seen any of the movies in their recommended list and if so whether they would agree with the rating provided. Secondly, we asked the users to read the details of one or two of the recommended movies which they hadn't watched before. This provided useful information as to whether users would be prepared to follow a recommendation and watch a movie based on the particular semantics of that movie e.g. type of genre, year of movie, actors appearing in a movie, directed by a certain director etc. Once users were presented with recommendations we asked them to firstly tell us (if they had previously seen the movie) whether they agreed with the predicted rating. The majority of users agreed with the predicted ratings to within a one or two ratings difference. However it was interesting to note that during the course of rating well-known or 'popular' movies there appeared to be two distinct sets of users emerging: one set who rated these movies highly and another set who tended to give these movies low ratings.

Influences on decisions to watch or purchase a movie	*Number of users(out of 15)*
Director	***6***
Actor	***7***
Certificate	***3***
Plot	***10***
Release Date	***7***
Title	***3***
Genre	***10***
Popularity	***7***
Availability	***3***
Cover images	***1***
Price (if purchasing)	***4***
External reviews/ user ratings	***8***
Already owned in a different format (if purchasing)	***1***
Language of movie	***1***

Table 2 Observational study findings

Whenever users were asked to look at the recommendations for movies which they had not previously seen, it was observed that the movies which they preferred had the same characteristics as the movies whose pages they had viewed while browsing. Some users were not prepared to consider these movies if they were not of the same type as their "normal" genre. Others would not watch these recommendations if they did not know the actors or directors.

We also asked users to tell us the contexts in which they browsed movie web-sites. Responses included the following: looking for information on new releases; checking out films someone had verbally recommended to them; in response to an email sent by the movie retailer (e.g. price reductions, new releases); purchasing for themselves or as a gift or simply "*just to browse*".

The main conclusions from the observational study were as follows

- Users were often not aware that they browsed movies based on the semantics of those movies until they were asked why they preferred a certain movie to another one. Furthermore, semantic information, such as actor, director or genre links, often guided the navigation of the site. When users were presented with recommendations from MovieLens, they tended to give high ratings to those movies whose semantics matched those movies they had preferred while browsing. In general different users displayed different semantic preferences. These observations provided the motivation for evaluating, through a user study, a semantically-based recommender system (Section 3.1 - GCM algorithm).

- The context of a user's visit appeared to depend on the end goal of the user e.g. 'just browsing' or 'purchasing a gift'. This provided the motivation for evaluating, through a user study, a semantic-based recommender which incorporates contextual information (Section 3.1 - GCMI algorithm).

- While rating MovieLens recommendations, two sets of users emerged: one group who liked 'popular' movies and one who did not. To test this behaviour further we asked users in our user study to rate a set of 'popular' movie recommendations (Section 3.4 - Top10).

3 User Study

3.1. Introduction

One of the goals of this study is to evaluate the use of additional semantic information on movies in the generation of recommendations. Thus, we employ a semantically-enhanced recommender algorithm called Generalized Cosine Max (GCM) as proposed in Anand, Kearney and Shapcott 2007. This algorithm is based on user-based collaborative filtering. However instead of simply measuring the similarity between two user ratings vectors, the algorithm incorporates an item ontology within the user vectors. For our study, individual page views of a movie web-site map to different concepts within the ontology. The similarity between items is calculated using information about a movie e.g. its actors, directors etc. and/or whether a user visited a particular actor or director web page. Furthermore, we want to evaluate whether the changing context of a user visit can be modeled using browsing tasks and whether by establishing the user's context we can improve upon the recommendations generated. For this we use an extension to the GCM algorithm (referred to as GCMI) which derives factors driving the observed user behaviour, referred to as impacts, and then incorporates these factors within the similarity computation.

Taking into account the feedback we received from the observational study we have decided to give users pre-defined browsing tasks (3.2) to carry out in the user study, rather than let them browse aimlessly, creating unnecessary noise. Furthermore, by giving users different scenarios we aim to simulate visits with different contexts and hence discover if this changing context has any affect on user behaviour and their personalised recommendations.

As a result of our observational study findings summarized in Table 2 we modify the GCM and GCMI algorithms to incorporate semantic information regarding actors, directors, year of creation, genre, certificate and title. We have decided not to include movie duration, format or region as in previous work (Kearney, Anand and Shapcott 2005).

3.2. Browsing tasks

Each of the user study browsing tasks is now discussed. The topology of the movie site used in the browsing tasks is discussed in 3.5.

Task One: *Starting from the home page, browse the movie site and find your top twenty movies of all time. Once complete provide a rating for each movie on a discrete scale of 1-5.*
The rationale behind this task is that the semantics which are most important to a user (as reflected in their top twenty movies) can be discovered and utilized in the recommendation process.

Task Two: *Starting from the home page, browse the movie site and find twenty movies you have previously watched. Rate each movie on a discrete scale of 1-5.*
The rationale of this task is aimed at establishing the features of movies which are the most important and the least important to a user (as reflected in a range of positive and negatively rated movies). The task is akin to that used by MovieLens where the user may choose any set of movies to rate within the registration process.

Task Three: *Starting from the home page, browse the movie site and find twenty movies which you would like to rent out to watch. You may include movies which you have or have not seen.*
The movies chosen in this scenario will not form part of the user behaviour utilized in 3.4 for recommendation generation. Since these are the twenty movies which the user most wants to see, intuitively it would seem prudent that the recommender system should generate recommendations for some of these movies. This scenario will therefore be used to test how good the generated recommendation sets are.

Task Four, Five and Six: *Using one of your top twenty movies as a starting point, browse the movie site and find ten other movies which you would like to find out more information on. You may include movies which you have or have not seen.*
In these tasks the user's top three movies were selected from the user's top twenty as a starting point. (If the second movie was very similar to the first, for example if

the user had given top ratings to "Star Wars 3" and "Star Wars 5", then the third and fourth movies would be chosen, etc.)

A user can navigate through the movie site using semantic links e.g. actor, genre etc. In these tasks a user will start at a favorite movie and then browse. It is hoped that from this starting point, a user will follow the semantic links that best describe his/her reasons for liking the movie that the search starts from. By using three different favorite movies the search space of the user in the movie site is increased. It is worth noting that the first three tasks are context-free: the user is not asked to group his/her favorite movie based on the context within which it was watched. On the other hand, in tasks 4 through to 6, we are attempting to simulate the changing context of a user's visit, by using a different movie as a starting point for each task.

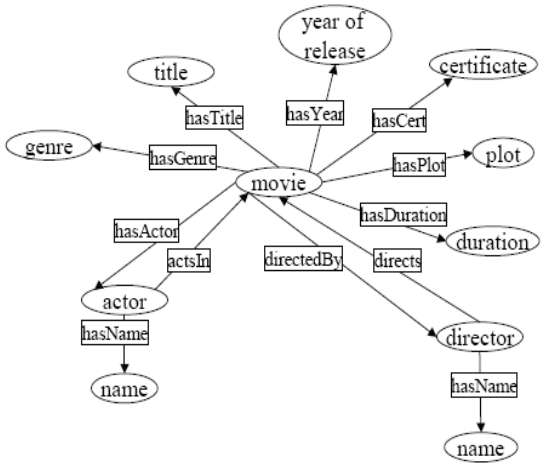

Figure 1 Movie Ontology

3.3 Architecture of user study

Data Acquisition To enable us to gather users' browsing behaviour in each of the tasks we utilize the Mozilla Firefox Eventlogger extension. This tool records the clickstream behaviour of the user during the six browsing tasks (3.2). Once processed, this information consists of all movie, actor and director pages which were browsed by each user, during each task. Users also provide ratings for the movies browsed within each browsing task and these are also added to this database to complete each user profile. Whenever an actor or director page is browsed by a user in the course of completing a particular task, the rating for that page is calculated as the average movie rating for that particular visit. Pages which do not correspond to an instance of the Movie, Actor or Director concepts within the movie ontology are ignored within the current evaluation. The browsing behaviour and the corresponding ratings form Dataset 1 in Figure 2 overleaf. In addition to the user behavior dataset, we also created a dataset of information describing the items which users could browse (Dataset 2 on Figure 2).

This knowledge base was generated by creating a web spider to crawl and parse the movie retailer's web site for additional semantic information about movies, actors and directors. For example, for each movie in the database, we have information relating to the plot, actors, director etc. We currently have information pertaining to 122,433 movies, 72,123 actors and 12,220 directors. This knowledge base conforms to the movie ontology shown in Figure 1.

Thirdly, to generate recommendations we required a further dataset of movie web site visitors with ratings that may have similar tastes to our users. Thanks to Netflix we were able to use their new publicly available dataset[2]. We decided to use the Netflix dataset because the original user behaviour dataset which was used in previous work (Anand, Kearney and Shapcott 2007), where no evaluation with users was carried out, only contained ratings for movies up to 2001. The new Netflix dataset, on the other hand, was collected between October, 1998 and December, 2005 and reflects the distribution of all movie ratings received during this period. The Netflix dataset contains movie rating files of over 100 million ratings from 480 thousand randomly-chosen, anonymous Netflix customers over 17 thousand movie titles. The date of each rating and the title and year of release for each movie are also provided. Previously, in the introduction, we stated that we wished to test the hypothesis that the context of a visit may be different from one user's visit to their next. To test this we grouped ratings provided by a Netflix visitor on a particular date together and assigned these to unique visits for that visitor.

As there are over 100 million ratings in the Netflix data, a training dataset was created where each visit contained a minimum of ten rated movies. This training data consisted of 15,777 visits which contained 1,010,960 ratings on 7,547 unique movies. The average number of movies rated per visit was 64 and the average rating over the entire training dataset was 3.65. Although the user behaviour dataset comes from the Netflix website it was decided not to use this site as the browsing site in the user trials. This is because the topology of this site does not offer a user the option to browse by movies starring particular actors or directed by particular directors.

Next, using the semantic information from Dataset 2, ontological profiles are created for both the users from the study and the Netflix users as per Anand, Kearney and Shapcott 2007.

Finally, recommendations are generated and presented to the users for evaluation using the recommendation techniques described in 3.4.

3.4. Recommendation generation

Four different algorithms are employed to generate recommendations for users in the study. For each of the algorithms (other than Top10) all visits by visitors within the training set are treated as separate user profiles to generate the neighborhood. A neighborhood size of fifty

[2] www.netflixprize.com

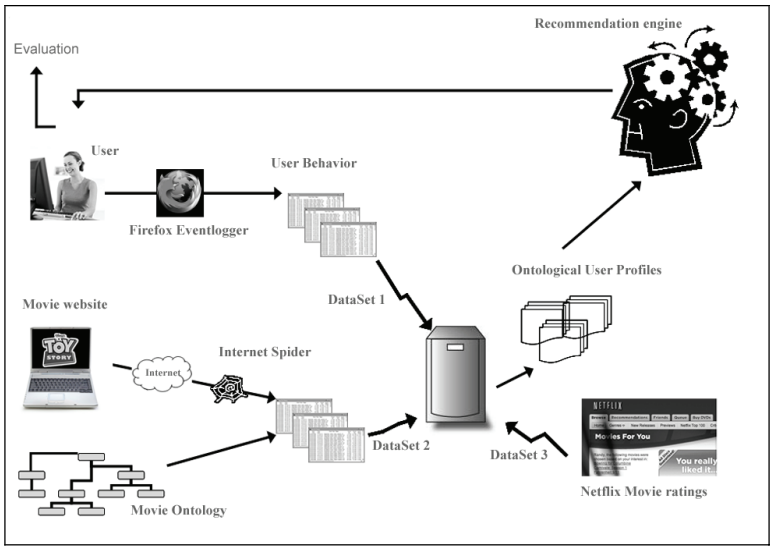

Figure 2 Architecture of User Study

has been chosen, based on previous empirical evaluation. Similarly each browsing task carried out by a user in the study is also treated as a separate user profile. In each of the three algorithms the predicted rating for items is calculated using the weighted average of the ratings of the items by visits within the neighborhood. Those used are:

- **Top10**: In this benchmark algorithm the ten most frequently rated items within the training data are selected and used as recommendations for all users. The predicted rating for these items will be calculated as the average rating for the items in the training data.

- **Visit Based Collaborative Filtering (VBCF)**: This algorithm will be used as a baseline aimed at measuring the efficacy of the similarity metrics used in the GCM and GCMI algorithms, with the neighborhood of visits selected using the cosine similarity metric.

- **Generalized Cosine Max (GCM)**: In this technique the (GCM) similarity metric is used to generate the neighborhood for each visit within the test data (Anand, Kearney and Shapcott 2007). This metric calculates user similarity by computing the similarity of the semantics of the items which they have rated.

- **Generalized Cosine Max using Impacts (GCMI)**: For each visit within the training and test data sets, impact values are calculated. These are factors which drive the observed user behaviour. During neighborhood formulation, the impact values for the target visit and the candidate neighbor visit are then used to generate the attribute weights used within the similarity calculation in the GCM similarity metric.

3.5. Movie website used in User Study

In browsing tasks 1, 2 and 3 (3.2) users were asked to browse a movie site starting from the hyperlinks provided on the movie site's home page. The information on a movie page conforms to the movie ontology as shown in Figure 1. From a movie page the user can navigate the site using actor, director, year of creation and movie format links. The topology of a movie page is shown in Figure 3. Links are highlighted in italics.

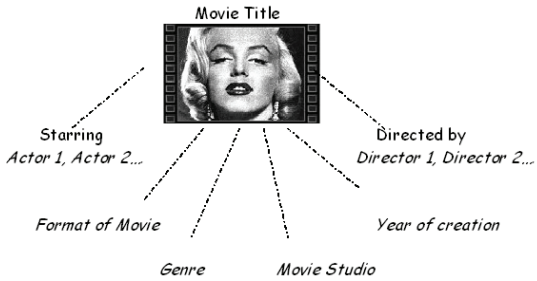

Figure 3 Topology of movie page

If a user chooses an Actor link e.g. "Tom Cruise" they will be directed to a URL which lists all movies which Tom Cruise acted in. The same pattern applies to director, genre, movie studio, year of creation and format.

4 Evaluation

4.1 User Evaluation

Five users took part in the user study which was intended to be a pilot for a larger experiment. Two of these users were academics (computer scientists) and would regard themselves as 'movie-buffs' and the other three were non-academics and occasional movie watchers.

Each user was asked to carry out the browsing tasks described in 3.2. For the pilot study we provided a quick orienteering session on the navigation of our chosen movie site. The aim of this session was to try to stop additional noise entering into the users' behavior through the use of 'haphazard' browsing.

Once each user completed the six browsing tasks the four techniques described in 3.4 were employed to generate recommendations. This generated sixteen sets of recommendations (three for each of the tasks other than Task 3 and one Top10 recommendation). As previously mentioned Task three was not used to generate a recommendation set, but was used for evaluation of the recommendation sets generated by the other tasks.

Each recommendation set was then presented to a user and they were asked to record whether they had watched (W) or not watched (NW) a movie previously. Next they were asked to rate each of the movies within a set on a scale of 1 – 5. Two ratings guidelines were provided as follows. If users had previously watched the movie the ratings they used were 5 - "Loved it", 4 – "Liked it", 3 – "Neither Liked nor Disliked It", 2 – "Did not like it" and 1 – "Hated it". If users had not watched the movie previously they were asked to follow the link to further information on the movie. After reading this information they had to decide what rating they would be prepared to give this movie based on the following scale: 5 – "Would definitely watch it", 4 – "Would probably watch it", 3 – "Not sure", 2 – "Would probably not watch it" and 1 – "Would definitely not watch it". Once users had completed their ratings for an entire set they were asked to give their overall verbal opinion on the usefulness of this set. Interestingly, users used the semantics of the movie to describe what they liked about a set, e.g ."*A good selection of comedy & action movies*". Finally they were asked to rank the sixteen sets of recommendations in order of preference and state why they preferred their top set. Although, we didn't use click-through as an evaluation metric, each user was asked to follow a link for further information on each movie, within a recommendation set and then rate it.

4.2 Evaluation Results

Table 3 shows the correlation between the ranking given to the recommendation sets and the average rating given to movies within the recommendation sets. As can be seen from the table, in three out of the five users, the correlation is lower; suggesting that the measure of desirability of a recommendation set, for these users, may be determined in a different way.

	Spearman's Rank Correlation Coefficient
User 1	0.65
User 2	0.93
User 3	0.69
User 4	0.82
User 5	0.66

Table 3 Correlation between Recommendation Set Ranking and Average Rating to Movies within the Recommendation Sets

Of the five users, Users 1 and 3 were avid movie fans (User 3 had not previously watched only 15% of the movies recommended to him across the 16 recommendation sets) while the other three could be described as "average" movie watchers. Table 4 shows the evaluation results obtained from each of the five users for recommendations generated using the Top10 algorithm. It is interesting to note that Users 4 and 5 assign the highest mean rating to the movies in the recommendation set generated using the Top10 algorithm while Users 2 and 5 ranked the recommendation set 2 out of the 16 sets presented to them. This suggests that "average" users are often happy to watch just the most popular movies. Interestingly, Users 1 and 3 rated this Top10 set 16th and 13th and also gave relatively low ratings to the recommendations produced. For User 3 the VBCF, GCM and GCMI algorithms each outperformed the Top10 algorithm and for User 1 both the VBCF and the GCM algorithms outperformed the Top10.

	User Ranking of Rec Set	Ranking by Mean Rating	Mean Rating (Top10)	MeanRating (VBCF, GCM, GCMI)
User 1	16	11	3.8	3.98, 4.04, 3.76
User 2	2	5	3.6	3.42, 3.58, 2.88
User 3	13	12	3.4	3.52, 3.7, 3.68
User 4	6	1	3.4	2.98, 3.04, 2.88
User 5	2	1	3.6	2.96, 3.12, 2.96

Table 4 Evaluation of Top10

We therefore hypothesize that although generating recommendations based on the top ten rated movies from the training data performed well for the "average" movie watcher, it cannot be relied upon for the more 'experienced' movie fan. Therefore, it would be useful if this type of user information could somehow be elicited from a user initially and fed in to the recommendation process.

Table 5 shows the Rankings and Mean Ratings assigned to the recommendation sets by each of the users. As can be seen from the table, based on the user rankings of recommendation sets, GCM outperforms the VBCF and GCMI algorithms in three out of the five tasks (and performs equally well as VBCF in Task 6), while GCMI outperforms VBCF and GCM in Task 4. Based on the Mean Rating, however, VBCF outperforms GCM and GCMI in Tasks 4 and 6. As for the appropriateness of these tasks to generate the seed set of ratings for the user, this appears to depend on which algorithm is being used. Task 6 appears to be best for VBCF, Task 2 for GCM and Task 4 for GCMI using User Rankings. Using the average rating to rank recommendation sets, Task 6 appears to perform best for VBCF, Task 1 performs best for GCM and Tasks 1 and 2 perform equally best for GCMI.

As mentioned previously, in Task 3 users were asked to provide twenty movies which they would like to see. Table 6 shows the results of this evaluation in terms of the number of recommendations found which match the users' movie selection in Task 3 and the average rank within the

Recommendation sets generated from	User 1			User 2			User 3			User 4			User 5			Average per algorithm per task		
	VBCF	GCM	GCMI	VBCF	GCM	GCMI	VBCF	GCM	GCMI	VBCF	GCM	GCMI	VBCF	GCM	GCMI	VBCF	GCM	GCMI
Task 1																		
User Ranking	6	1	13	5	3	15	9	5	2	10	11	13	10	5	9	8.00	5.00	10.40
Mean Ratings	3.8	4.6	3.9	3.6	4	2.9	3.6	3.7	4	2.8	2.9	3	3.4	3.2	3.2	3.44	3.68	3.40
Task 2																		
User Ranking	3	9	14	14	7	12	3	1	4	16	2	8	16	1	7	10.40	4.00	9.00
Mean Ratings	4.5	3.9	3.6	3.3	3.6	3.1	3.8	4.3	4.1	2.8	3.2	3.2	2.3	3.3	3	3.34	3.66	3.40
Task 4																		
User Ranking	8	11	7	8	10	9	10	11	6	3	1	5	15	11	6	8.80	8.80	6.60
Mean Ratings	3.9	3.9	4.1	3.4	3.1	3.4	3.7	3.4	3.4	3.2	3.4	2.9	3.1	2.9	2.9	3.46	3.34	3.34
Task 5																		
User Ranking	12	2	15	11	6	13	14	7	15	7	9	15	12	8	14	11.20	6.40	14.40
Mean Ratings	3.6	4.2	3.3	3	3.6	2.8	3.3	3.6	3.1	3	2.9	2.6	2.8	2.9	3.1	3.14	3.44	2.98
Task 6																		
User Ranking	10	4	5	1	4	16	12	8	16	4	12	14	4	3	13	6.20	6.20	12.80
Mean Ratings	4.1	3.6	3.9	3.8	3.6	2.2	3.2	3.5	3.8	3.1	2.8	2.7	3.2	3.3	2.6	3.48	3.36	3.04
Average User Ranking	7.80	5.40	10.80	7.80	6.00	13.00	9.60	6.40	8.60	8.00	7.00	11.00	11.40	5.60	9.80	8.92	6.08	10.64
Average Mean Rating	3.98	4.04	3.76	3.42	3.58	2.88	3.52	3.70	3.68	2.98	3.04	2.88	2.96	3.12	2.96	3.37	3.50	3.23

Table 5 Rankings and mean ratings of recommendation sets by users

recommendation list generated by the algorithms. Interestingly the GCMI algorithm that did not perform well within the user trial seems to significantly outperform the other algorithms for this metric.

	Number of Matching Recommendations	Average Rank
VBCF	15	53
GCM	16	49
GCMI	33	36

Table 6: Evaluation against Task 3 movie selection

5 Related Work

We compare our approach to related work. The MovieLens recommender has provided a useful platform for both interface design and algorithm evaluation by users of the movie domain (O'Connor et al. 2001, Herlocker, Konstan and Riedl 2000, Sen et al. 2006). However, unlike this study, no evaluation has been carried out which considers the importance of the underlying semantics of the movie domain or the importance of establishing the context behind a user's visit.

Geyer-Schulz, Hahsler and Jahn (2001) evaluated their recommender by randomly selected 300 lists of recommended items and asking researchers if they would recommend the later items to a person interested in the first one. In this study the focus is more on discovering whether association between items exists, rather than a deeper understanding of each item within a domain.

For the Quickstep recommender system, two user trials were conducted using ontological profiles (Middleton, Shadbolt and Roure 2004). Although we also evaluate ontological profiles the focus of our study differs as we don't assume the availability of external ontologies, which while applicable in a research paper domain, cannot however be assumed in a general e-tailer domain scenario such as movies. Furthermore the context of a user's visit is not considered.

Haase et al. (2005) evaluate semantic user profiles from usage and content information to provide personalized access to bibliographic information on a Peer-to-Peer bibliographic network. Similar to the GCMI approach, they do evaluate the impact of different concepts on the users' preferences. However, rather than these weights being provided by the user as in this study, we evaluate weights which have been calculated from observing the users' behaviour.

6 Conclusions and Future Work

In this paper we have presented the results of a pilot user study into the effectiveness of collaborative recommender systems that incorporate item semantics within the recommendation generation process. Satisfaction with a set of recommendations was often defined in terms of overall semantic attributes of the set, suggesting that the use of semantics within recommendation generation should provide improved user satisfaction. This was shown to be the case in our user study where the GCM algorithm outperformed the VBCF algorithm. Another interesting finding of the study was the fact that the collaborative filtering algorithms outperformed the simple Top10 algorithm only for "movie buffs".

A somewhat surprising result was the fact that the contextually enhanced recommender algorithm (GCMI) appeared to be unhelpful. However, given that this algorithm has been shown to improve precision and recall in previous studies (Anand, Kearney and Shapcott 2007) and also in the evaluation against Task 3 items, it would appear that the algorithm merits further investigation. One possible reason for the algorithm underperforming may be the artificial nature of Tasks four to six used in this study may not reflect context adequately. Another possibility is that GCMI may have performed better if, during the evaluation, the users had been reminded of what movies they browsed in each task, to remind them of the context of that task. On the whole Tasks 1 and 2 perform well as seed ratings, suggesting that information on a user's top movies or the collection of positive movie ratings along with negative movie ratings from a user may provide the best

results in terms of generating desirable recommendation sets. However, for the VBCF there is evidence that more directed approaches such as those in Tasks 4 though to 6 may actually provide better user profiles.

Due to the limited number of participants in the user study the results presented here are not statistically significant. However, it is interesting to note that two user groups with similar behaviour appear to have been identified from the study – the so-called 'movie-buff' and the 'non-movie buff'. We therefore plan to carry out a further study with more users with representation from both of these groups to establish the validity of our findings. We will also be examining the GCMI algorithm more carefully to see if we can design more realistic tasks that would favor the incorporation of impact values. Previous work with hidden test data indicated that it was capable of generating more of the hidden movies than other algorithms. On the whole we found that the correlation between rankings for user preference for the recommendation sets and the mean rating of a set appear to be dependent on the user. We therefore plan to investigate more closely the correlation between users' ratings and how they rank recommendation sets in the further study.

Finally, although we evaluate the movie domain in this research, we believe that this study could be generalized to other domains.

Acknowledgements

This research was partly funded by the Department of Employment and Learning Northern Ireland.

References

Adomavicius, G. and Tuzhilin, A. 2005. Towards 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).

Anand, S.S., Kearney, P. and Shapcott, C. M. Generating Semantically Enriched User Profiles for Web Personalization, 2007. *ACM TOIT*: 7(4). Forthcoming..

Anand, S. S. and Mobasher, B. 2005. *Intelligent techniques in Web Personalization*. In Intelligent Techniques in Web Personalization, B. Mobasher and S. S. Anand, Eds. LNAI 3169. Springer-Verlag, 1-37.

Belkin, N. and Croft, B. 1992. Information Filtering and Information Retrieval. Communications of the *ACM*: 35(12):29-37.

Borchers, A., Herlocker, J., Konstan, J. and Riedl, J. 1998. Ganging Up on Information Overload. *IEEE Computer*:106-108.

Burke, R. 2002. Hybrid recommender systems: Survey and experiments. *User Modeling and User Adapted Interaction* 12(4): 331-370.

Cho, Y. and Kim, J. 2004. Application of web usage mining and product taxonomy to collaborative recommendations in e-commerce. *Expert Systems with Applications* 26 (2): 233-246.

Geyer-Schulz, A., Hahsler, M. and Jahn, M. 2001. A customer purchase incidence model applied to recommender systems. In *Proceedings of the ACM WebKDD Workshop on Mining Web Log Data Across All Customer Touch Points*. San Francisco, CA.

D. Gupta, M. Digiovanni, H. Narita ,K. Goldberg. 1999. Jester 2.0: Evaluation of a New Linear Time Collaborative Filtering Algorithm In *Proceedings of the 22nd ACM SIGIR conference*, 291 – 292.

Haase, H., Hotho, A., Schmidt-Thieme, L. Sure, Y. 2005. Collaborative and Usage-Driven Evolution of Personal Ontologies. In *Proceedings of ESWC*: 486-499.

Herlocker, J., Konstan, J., and Riedl, J. 2000. Explaining Collaborative Filtering Recommendations. In *Proceedings of the ACM Conference on Computer-Supported Cooperative Work.* Philadelphia Penn.

Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM TOIS*: 22(1), 5-53.

Kearney, P., Anand, S. S., and Shapcott, C. M. 2005. Employing a Domain Ontology to Gain Insights into User Behaviour. In Working Notes of the IJCAI Workshop on Intelligent Techniques for Web Personalization, 25-32. Edinburgh, Scotland.

Middleton, S. E., Shadbolt, N. R., and Roure, D. C. D. 2004. Ontological User Profiling in Recommender Systems. *ACM TOIS*: 22 (1) 54-88.

Mobasher, B., Jin, X., and Zhou, Y. 2004. Semantically Enhanced Collaborative Filtering on the Web. *In Web Mining: From Web to Semantic Web*. LNAI 3209. Springer-Verlag.

O'Connor, M., Cosley, D., Konstan, J. A., & Riedl, J. 2001. PolyLens: A Recommender System for Groups of Users. In *Proceedings of ECSCW 2001*,199-218.

Preece, J., Rogers, Y., Sharp, H., Benyon, D., Holland, S. & Carey, T. 1994. *Human-Computer Interaction.* Reading MA: Addison-Wesley

Sen, S., Lam, S.K., Cosley, D., Rashid, A.M., Frankowski, D., Harper, F., Osterhouse,J. and Riedl, J. 2006. Tagging, community, vocabulary, evolution. In *Proceedings of CSCW 2006*.

Ziegler, C., McNee, S. M., Konstan, J. A., and Lausen, G. 2005. Improving Recommendation Lists through Topic Diversification. In Proceedings of the 14th international conference on World Wide Web. 22-32.