Investigating Serendipity in Recommender Systems Based on Real User Feedback

Denis Kotkov
Faculty of Information Technology,
University of Jyvaskyla, Finland
deigkotk@student.jyu.fi

Joseph A. Konstan
University of Minnesota
Minneapolis, MN, USA
konstan@umn.edu

Qian Zhao
University of Minnesota
Minneapolis, MN, USA
zhaox331@umn.edu

Jari Veijalainen
Faculty of Information Technology,
University of Jyvaskyla, Finland
jari.veijalainen@jyu.fi

ABSTRACT

Over the past several years, research in recommender systems has emphasized the importance of serendipity, but there is still no consensus on the definition of this concept and whether serendipitous items should be recommended is still not a well-addressed question. According to the most common definition, serendipity consists of three components: relevance, novelty and unexpectedness, where each component has multiple variations. In this paper, we looked at eight different definitions of serendipity and asked users how they perceived them in the context of movie recommendations. We surveyed 475 users of the movie recommender system, MovieLens regarding 2146 movies in total and compared those definitions of serendipity based on user responses. We found that most kinds of serendipity and all the variations of serendipity components broaden user preferences, but one variation of unexpectedness hurts user satisfaction. We found effective features for detecting serendipitous movies according to definitions that do not include this variation of unexpectedness. We also found that different variations of unexpectedness and different kinds of serendipity have different effects on preference broadening and user satisfaction. Among movies users rate in our system, up to 8.5% are serendipitous according to at least one definition of serendipity, while among recommendations that users receive and follow in our system, this ratio is up to 69%.

CCS CONCEPTS

•Information systems → Recommender systems; *Personalization;*

KEYWORDS

recommender systems; serendipity; relevance; novelty; unexpectedness

SAC 2018, Pau, France
© 2018 ACM. 978-1-4503-5191-1/18/04…$15.00
DOI: 10.1145/3167132.3167276

ACM Reference format:
Denis Kotkov, Joseph A. Konstan, Qian Zhao, and Jari Veijalainen. 2018. Investigating Serendipity in Recommender Systems Based on Real User Feedback. In *Proceedings of SAC 2018: Symposium on Applied Computing , Pau, France, April 9–13, 2018 (SAC 2018),* 10 pages.
DOI: 10.1145/3167132.3167276

1 INTRODUCTION

Recommender systems are designed to help users find interesting items, when the number of these items is overwhelming [24]. In this paper, the term *item* refers to a piece of information, which is a reference to an object, such as a good, service or process that a recommender system suggests to the user [17]. An item can refer to any object, such as a movie, song or book.

Traditionally, recommendation algorithms have been optimized for accuracy [15], which indicates the predictive power of these algorithms. However, recently the focus of the recommender systems community started shifting towards factors beyond accuracy [15], as accuracy alone does not always result in user satisfaction. One of the factors of recommender systems beyond accuracy is serendipity [15].

According to the dictionary, serendipity is "the faculty of making fortunate discoveries by accident"[1]. The term serendipity was first introduced in the context of recommender systems in early 2000s [10]. Many researchers employed their definitions of this concept, but there is no consensus on the definition of serendipity yet [17, 19]. The most common definitions of the concept include three components: relevance, novelty and unexpectedness [13, 19, 21], while these components have multiple definitions [19].

It is unclear whether serendipitous items should be recommended to users. According to most claims from the literature on serendipity in recommender systems, there are two main reasons for collaborative recommender systems [8] to suggest serendipitous items: they broaden user preferences [10, 29, 30] and increase user satisfaction [1, 20, 22, 29]. However, the studies showing that serendipitous items are in any way better than relevant non-serendipitous ones are very limited and often have a small number of samples [26, 27, 29].

Novelty and unexpectedness have multiple definitions [16], but it is unclear whether different variations of the same component

[1] http://www.thefreedictionary.com/serendipity

have different effects on our metrics. In this paper, the term *metrics* refers to preference broadening and user satisfaction.

Researchers often indicate that serendipitous items are very rare [4, 17]. However, it is unclear exactly how rare these items are, as it might not be worth of suggesting them due to their rareness and a high risk of suggesting irrelevant items, while optimizing for serendipity [19].

This paper presents the first study that looks across multiple definitions of serendipity. It compares these definitions and their components in terms of their value for a user in a user study. We employ the most common definition of the concept, which requires serendipity to include three components: relevance, novelty and unexpectedness, where each component has multiple definitions resulting in eight definitions of serendipity. We detect the most important features for predicting serendipitous items and estimate the ratio of these items among items rated in a typical collaborative-filtering-based recommender system [8]. We conduct this study in the online movie recommender system, MovieLens[2], where we ask users retrospectively about movies they have rated. In this paper, we address the following research questions (RQs):

RQ1. What are the effects of variations of serendipity components on broadening user preferences and user satisfaction?

RQ2. What are the effects of different kinds of serendipity on broadening user preferences and user satisfaction?

RQ3. What are the effective features for detecting serendipitous movies? What are the value ranges of these features for typical serendipitous movies?

RQ4. How rare are serendipitous movies among movies rated by the users in a typical collaborative-filtering-based recommender system? To what extent does this kind of system help users find these movies?

Most serendipity-oriented algorithms are evaluated based on publicly available datasets that lack user feedback regarding serendipitous items. To label certain items serendipitous to users, researchers tend to make assumptions regarding serendipity, such as serendipitous items are unpopular [20, 22, 30] or dissimilar to items users rated in the past [1]. These assumptions might not correspond to real life scenarios. We therefore publish our collected dataset to allow other researchers conduct experiments related to serendipity in recommender systems[3]. This paper has the following contributions:

- We conduct literature review and operationalize common definitions of serendipity.
- We compare different definitions of serendipity and their components in terms of preference broadening and user satisfaction.
- We find a subset of features that are effective for detecting movies that are serendipitous according to certain definitions, which might be useful for suggesting these these serendipitous movies in an off-line evaluation of recommendation algorithms.
- We estimate the ratio of serendipitous movies, which might help to decide whether it is worth optimizing for serendipity.
- We publish the first dataset that includes user feedback regarding serendipitous movies to allow other researchers conduct their experiments.

In this study, we surveyed 475 users and found that most kinds of serendipity and all the variations of serendipity components broaden user preferences, but one variation of unexpectedness hurts user satisfaction. We found effective features for detecting movies that are serendipitous according to definitions that do not include this variation of unexpectedness. These features are predicted rating, popularity, content-based and collaborative similarity to movies users watched in the past. We also found that different variations of unexpectedness and different kinds of serendipity have different effects on preference broadening and user satisfaction. Among movies users rate in a typical collaborative-filtering-based system, up to 8.5% are serendipitous according to at least one definition of serendipity, while among recommendations that users receive and follow in the system, this ratio is up to 69%. We only provide an upper bound estimation due to the bias of our dataset (we selected relatively unpopular movies).

The paper is organized as follows: in section 2, we briefly review related work. Section 3 describes our survey and the method to invite users to our study. Section 4 describes the dataset we collected. In section 5, we analyze the collected data and answer our research questions. In section 6, we discuss the results, while in section 7 we discuss limitations of our study and future work. Finally, we conclude in section 8.

2 RELATED WORK

Many authors focused on different aspects of serendipity and used different definitions of the concept. In this section, we focus on (1) the value of serendipitous items for users due to the objective of our research, (2) definitions of serendipity, as this is the key concept of our research and (3) inquiring for serendipity, as we conduct a survey where we ask users to indicate serendipitous movies.

2.1 Why Serendipitous Items?

Most previous studies on this topic indicate three reasons to recommend serendipitous items. Researchers have claimed that serendipitous items help overcome the overspecialization problem (for content-based filtering algorithms) [1, 13], broaden user preferences [10, 29, 30] and increase user satisfaction [1, 20, 22, 29].

However, studies that provide evidence for the benefit of recommending serendipitous items are very limited. The only study we found that measured the benefit of serendipitous recommendations was conducted by Zhang at el. [29]. In the study, 21 users were offered recommendations from serendipity-oriented and accuracy-oriented algorithms. Although users gave lower ratings to recommendations provided by the serendipity-oriented algorithm than those provided by the accuracy-oriented algorithm, the majority of users preferred using the serendipity-oriented one [29].

To the best of our knowledge, there are no studies that compare items corresponding to different definitions of serendipity in terms of their value for users. In this paper, we compare different definitions of serendipity in terms of preference broadening and user satisfaction. We do not consider the overspecialization problem, as in MovieLens, users mostly receive recommendations generated by

[2] movielens.org
[3] The dataset is available on the GroupLens website: https://grouplens.org/datasets/movielens/

collaborative filtering algorithms [7], while the overspecialization problem is more prominent for content-based filtering algorithms [13].

2.2 The Definitions of Serendipity

There is no consensus on the definition of serendipity in recommender systems [17, 19]. However, most authors indicate that serendipitous items must be relevant, novel and unexpected to a user [17]. An item is relevant to a user if the user expresses or will express their preference for the item in the future by liking or consuming the item depending on the application scenario [18]. Novelty of an item to a user depends on how familiar the user is with the item. An item can be novel to a user in different ways:

(1) The user has never heard about the item [16].
(2) The user has heard about the item, but has not consumed it.
(3) The user has consumed the item and forgot about it [16].

Studies on serendipity in recommender systems often neglect the definition of unexpectedness. We present a number of definitions corresponding to the component. An item can be unexpected to the user if:

(1) The user does not expect this item to be relevant to them [1].
(2) The user does not expect this item to be recommended to them.
(3) The user would not have found this item on their own [1, 9–11, 27].
(4) The item is significantly dissimilar to items the user usually consumes [14, 19, 29].
(5) The user does not expect to find this item, as the user is looking for other kinds of items [1].

In this paper, we investigate serendipity according to different definitions: serendipitous items are relevant, novel and unexpected, where unexpectedness corresponds to definitions 1–4 and novelty corresponds to all the definitions listed above (we merged definitions 1 and 3 together). We do not consider definition 5 of unexpectedness, as many users in MovieLens do not normally look for particular kinds of movies. Furthermore, if they know what they are looking for, they are unlikely to remember what kinds of movies they were looking for after they have watched the movie they found.

2.3 Inquiry About Serendipity

There are two ways of inquiring about serendipity in surveys: posing questions concerning serendipity while viewing it as atomic, or exposing its components and posing suitable questions concerning them. The former way of inquiring requires less effort from a user and simplifies the analysis of user answers. However, asking one question does not allow to investigate components of serendipity and is likely to be confusing for users due to the complexity of the concept [25]. For example, Said et al. compared results of collaborative filtering algorithms in terms of serendipity in an online experiment [25]. The authors directly asked users whether they found recommendations serendipitous and received statistically insignificant results in terms of serendipity when they compared performance of the algorithms. The authors noted that this insignificance was caused by the complexity of the concept especially for non-native speakers [25].

Inquiring about each component of serendipity requires the users to answer several questions, where one or more questions measure one concept at a time. For example, Zhang et al. considered an item serendipitous to a user, when that user gave an item a high rating and indicated that this item was novel and unexpected to them [29]. Although this way of inquiring about serendipity is more demanding for users, it allows to investigate each component of serendipity and measure serendipity more precisely than asking just the one question.

It is also possible to use implicit user feedback on items to assess serendipity. For example, de Gemmis et al. analyzed facial expressions of users to detect the movies that were serendipitous to these users [6].

In this paper, we conduct a survey, where we inquire about serendipity by asking users one question per component of serendipity according to each definition employed in this research, as our goal of investigating each definition of serendipity requires precise assessment of the concept and its components.

3 THE SURVEY DESIGN

The main functionality of MovieLens allows users to rate movies they watched on the scale from 0.5 to 5 stars with the granularity of 0.5 star and receive recommendations generated based on the ratings. MovieLens does not allow users to indicate how long ago they had watched a particular movie. Users might rate a movie in a while after they had watched it. MovieLens also allows users to perform other actions, such as adding a movie to the list of movies to watch (a watch list), assigning keywords (tags) to movies, and adding new movies to the system.

The ideal way to measure serendipity in a movie domain would be to inquire a user about novelty and unexpectedness before the user has watched the movie and inquire the user about the relevance of this movie afterwards. MovieLens allows us to implement this experimental setting by conducting two surveys: the first one, when a user adds movies to their watch list and the second one, when the user rates movies from their watch list. However, this setting has two main disadvantages: (a) users mostly add movies they expect to enjoy watching to their watch lists, and (b) only a few users use the functionality of adding movies into watch lists and even fewer users rate movies from their watch lists. We therefore decided to ask users about their experience retrospectively.

We invited users via emails to complete an online survey regarding movies they rated during the last three months before the experiment. We chose three months, because it is likely that users still remember their experience of rating those movies when the users take our survey. Our inclusion criteria for users was as follows: we selected users who rated at least five movies with a rating of at least 3.5 stars from December 30, 2016 till March 30, 2017 (the experiment started on April 1, 2017) and at least one month after their registrations (for users who joined MovieLens after November 30, 2016). We assumed that users rate movies that they watched before the registration during the first month after their registration. In our survey, we picked five movies rated during the three months before the experiment and asked users to answer five questions and rate forty statements about the five movies we picked for each user (one question and eight statements per movie). We picked the least popular movies (i.e. those with the smallest number of ratings in the system) among movies users rated during that period of time. We expected that users discovered these movies in our system, as users are likely to hear about popular movies from other sources, such as friends, family and TV.

We emailed 2305 users who met our inclusion criteria and received a response from 522 users, but only 475 users rated all the statements and answered the question about at least one movie. In total, these users rated all the statements and answered all the questions about 2166 movies.

Table 1 demonstrates statements we asked users to rate. Serendipity components that correspond to the statements and the definitions of serendipity. We asked each user to rate eight statements using the following scale: "strongly agree", "agree", "neither agree nor disagree", "disagree", "strongly disagree", "don't remember". Each definition of serendipity consists of three components: relevance, novelty and unexpectedness. As we only asked users about movies they rated with at least 3.5 stars, we assumed that all the movies we asked users about are relevant to these users. We picked four definitions of unexpectedness and two definitions of novelty:

- Unexpectedness to be relevant (unexp_rel) corresponds to the original definition of unexpectedness 1 from the literature review section (section 2.2).
- Unexpectedness to be found (unexp_find) corresponds to the original definition of unexpectedness 3.
- Implicit unexpectedness (unexp_imp) corresponds to the original definition of unexpectedness 4.
- Unexpectedness to be recommended (unexp_rec) corresponds to the original definition of unexpectedness 2.
- Strict novelty (s_nov) corresponds to the original definitions of novelty 1 and 3.
- Motivationally novelty (m_nov) corresponds to the original definition of novelty 2.

This resulted in eight sets of serendipitous movies. For example, we considered a movie motivationally serendipitous (implicit) if a user rated statements 2 and 5 (m_nov and unexp_imp) with replies "strongly agree" or "agree". One movie can belong to several definitions simultaneously.

In this paper, the term *movie* refers to a user-movie pair. For example, a relevant movie corresponds to a user-movie pair, where the user considers the movie relevant, while other users might not consider this movie relevant.

4 SUMMARY STATISTICS OF THE DATASET

Figure 1 demonstrates the distribution of the answers to the question of how long ago users watched movies we picked for the survey. Users watched around 60% of the movies we asked them about less than 6 months before the survey and therefore it is likely that they still remember their watching experience for the movies. We removed movies that users indicated they did not watch (20 movies or 1%) from our dataset.

Figure 2 demonstrates distributions of the user responses. Users indicated that they were glad they watched the majority of movies

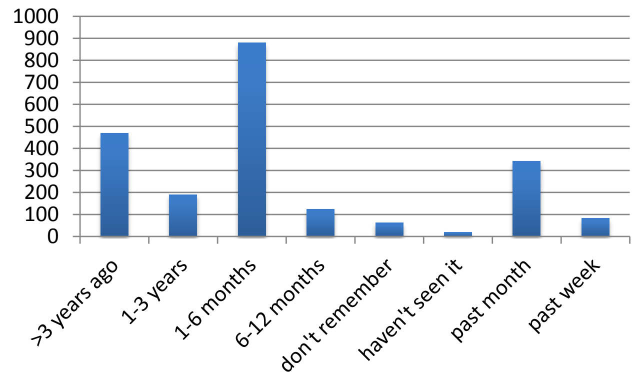

Figure 1: Distribution of answers to the question "When did you watch this movie for the first time?"

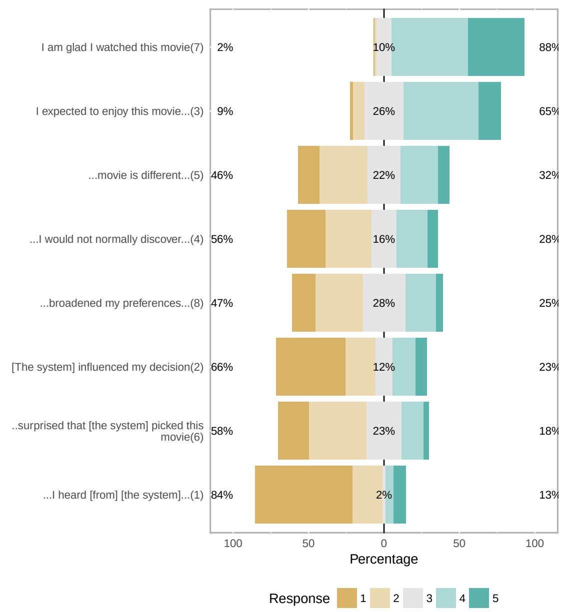

Figure 2: Distributions of answers ("1" - strongly diagree, "2" - disagree, "3" - neither agree nor disagree, "4" - agree, "5" - strongly agree)

we asked them about, which might have resulted from our inclusion criteria (we picked movies users rated at least 3.5 stars in MovieLens).

Table 2 demonstrates the numbers of movies that are serendipitous according to the different definitions along with all the movies we picked for the survey. The sets of different kinds of serendipitous movies overlap. For example, 48 m_ser_imp movies are at the same time m_ser_rec.

5 ANALYSIS

In this section, we explain how we analyzed the collected data set. All the movies we picked are considered relevant by users due to

Table 1: Statements 1-6 correspond to components of serendipity, statements 7 and 8 correspond to our metrics and "+" indicates inclusion of a component to the definition of serendipity, i.e. the user checks "agree" or "strongly agree" to the corresponding statement, except for Statement 3 where inclusion means checking "disagree", "strongly disagree" or "neither agree nor disagree".

Statement #	1	2	3	4	5	6	7	8
Statement	The first time I heard of this movie was when MovieLens suggested it to me.	MovieLens influenced my decision to watch this movie.	I expected to enjoy this movie before watching it for the first time.	This is the type of movie I would not normally discover on my own; I need a recommender system like MovieLens to find movies like this one.	This movie is different (e.g., in style, genre, topic) from the movies I usually watch.	I was (or, would have been) surprised that MovieLens picked this movie to recommend to me.	I am glad I watched this movie.	Watching this movie broadened my preferences. Now I am interested in a wider selection of movies.
Name	s_nov	m_nov	unexp_rel	unexp_find	unexp_imp	unexp_rec	Satisfaction	Preference broadening
Description	A novelty component (strict novelty)	A novelty component (motivational novelty)	An unexpectedness component (unexpectedness (relevance))	An unexpectedness component (unexpectedness (find))	An unexpectedness component (unexpectedness (implicit))	An unexpectedness component (unexpectedness (recommend))	Our satisfaction metric	Our preference broadening metric
s_ser_rel	+		+					
s_ser_find	+			+				
s_ser_imp	+				+			
s_ser_rec	+					+		
m_ser_rel		+	+					
m_ser_find		+		+				
m_ser_imp		+			+			
m_ser_rec		+				+		

our inclusion criteria. In the following text, we omit indicating that the movies are relevant for brevity.

We employed a cumulative link mixed-effect regression model [5]. We used this model to predict a dependent ordinal variable Y with independent binary variables $x_1, x_2, ..., x_n$. Consider an observation from the kth user. The model fits probability of the dependent variable to fall in $j = 1, 2, ..., J$ categories as follows:

$$log\left(\frac{P(Y \leqslant j)}{1 - P(Y \leqslant j)}\right) = \alpha_j + \beta_1 x_1 + ... + \beta_n x_n + u_k, \qquad (1)$$

where $P(Y \leqslant j)$ is a cumulative probability that $Y \leqslant j$, while $\beta_1, \beta_2, ..., \beta_n$ are the coefficients of the model, where n depends on the number of independent variables included in the model. Parameter α_j corresponds to the intercept for category j, while u_k is a random intercept for the kth user and $u_k \sim Gaussian(0, \sigma)$ where σ is an additional parameter describing the dispersion of the random intercept effect. See [2] for more details on this analytical model for ordinal response data. We use the R *ordinal* package (a standard implementation of the model) to conduct the analysis [5].

In our analysis, the dependent variables correspond to the likert-scale responses users gave to the statements 7 and 8 (preference broadening or user satisfaction, separately). They take values from 1 to 5. Our independent variables correspond to whether a movie satisfies a particular definition of serendipity or a variation of serendipity component.

We conducted statistical tests for each coefficient of the regression models, where the null hypothesis was that the coefficient equals zero meaning that changes in the independent variable were not associated with changes in the dependent variable. To control false discoveries, we used Bonferroni correction procedure, which adjusted our critical p-value from 0.05 to 0.0005 [12].

5.1 Effects of the Serendipity Components

To answer RQ1 and investigate the effects of variations of serendipity components separately, we ran twelve cumulative link mixed-effect regression models, two models per component variation. In each model, we predicted the dependent variable (user responses to the preference broadening or user satisfaction question) with a binary independent variable, i.e. whether a movie belongs to a particular variation of the component. The results are summarized in Table 3. It shows that:

(1) Movies that are novel according to either definition of novelty have a positive effect on preference broadening compared with the corresponding non-novel movies.
(2) Movies that are unexpected according to each definition of unexpectedness have a positive effect on preference broadening compared with the corresponding non-unexpected movies.
(3) unexp_rel movies have a negative effect on user satisfaction compared with the non unexp_rel movies.

Table 2: General characteristics of the dataset. Strict serendipity is the union of all kinds of strict serendipity (movies corresponding to at least one definition of serendipity that requires strict novelty (s_nov)), motivational serendipity is the union of all kinds of motivational serendipity, and serendipity is the union of all kinds of serendipity.

Concept	Movies	Users
All	2146	475
Strictly serendipitous (relevant) (s_ser_rel)	77	61
Strictly serendipitous (find) (s_ser_find)	181	119
Strictly serendipitous (implicit) (s_ser_imp)	115	80
Strictly serendipitous (recommend) (s_ser_rec)	63	50
Strictly serendipitous	205	131
Motivationally serendipitous (relevant) (m_ser_rel)	91	64
Motivationally serendipitous (find) (m_ser_find)	163	101
Motivationally serendipitous (implicit) (m_ser_imp)	128	88
Motivationally serendipitous (recommend) (m_ser_rec)	71	49
Motivationally serendipitous	218	122
Serendipitous	302	173

Table 3: The fixed effects (coefficients) of the twelve cumulative link mixed-effect regression models. Each cell corresponds to a coefficient of an ordinal regression with a single independent variable. Dependent variables are our metrics (broadening or satisfaction), while independent variables are variations of serendipity components (metric ∼ component). Significance codes: "*" < 0.0005.

Component	Broadening	Satisfaction
s_nov	0.7412*	0.1259
m_nov	0.7827*	0.307
unexp_rel	0.5972*	-0.9133*
unexp_find	2.1161*	0.1934
unexp_imp	1.8698*	-0.03029
unexp_rec	1.0889*	-0.451

All variations of novelty and unexpectedness broaden user preferences (observations 1 and 2), but unexpectedness (relevant) (movies that are unexpected to be relevant) hurts user satisfaction (observation 3).

Next, we conducted direct comparisons between the variations of novelty and unexpectedness. For the unexpectedness component, we ran twelve cumulative link mixed-effect regression models, two models per comparison. We ran each model on a subset of the collected dataset where we included only observations belonging to the two variations that were being compared. In the dataset, we repeated observations belonging to the two variations simultaneously. Table 4 summarizes the results for the variations of the unexpectedness component. It shows that:

Table 4: The fixed effects (coefficients) of the twelve cumulative link mixed-effect regression models. Each cell corresponds to a coefficient of an ordinal regression with a single independent variable run on a dataset consisting of instances belonging to variations of unexpectedness indicated in the left column and top row. Dependent variables are our metrics (broadening or satisfaction), while independent variables are variations of unexpectedness indicated in the left column (metric ∼ component). Significance codes: "*" < 0.0005

Broadening			
Component	unexp_rel	unexp_find	unexp_imp
unexp_find	0.8206*		
unexp_imp	0.6325*	-0.1343	
unexp_rec	0.4079	-0.3932	-0.1857
Satisfaction			
Component	unexp_rel	unexp_find	unexp_imp
unexp_find	0.6373*		
unexp_imp	0.4634*	-0.1463*	
unexp_rec	0.1436	-0.4529	-0.3270

(1) unexp_find and unexp_imp movies have positive effects on preference broadening and user satisfaction, when compared with unexp_rel movies.
(2) unexp_find movies have a positive effect on user satisfaction, when compared with unexp_imp movies.

Variations of unexpectedness components turned out to differ in terms of our metrics. Unexpectedness (relevant) broadens user preferences less and results in a lower user satisfaction than unexpectedness (find and implicit) (observation 1), while unexpectedness (find) outperforms unexpectedness (implicit) in terms of user satisfaction (observation 2).

We omitted the results for the variations of novelty, because we did not find any statistically significant results in comparisons between strict and motivational novelty in terms of preference broadening and user satisfaction.

5.2 Effects of Serendipity

To address RQ2, we ran sixteen cumulative link mixed-effect regression models, two models per serendipity definition. In each model, the dependent variable corresponds to the metric (preference broadening or satisfaction), while the independent variable is a binary variable, which equals true if the movie is serendipitous according to a particular definition of serendipity and false otherwise.

Table 5 summarizes the results and shows that movies that are serendipitous according to seven definitions of serendipity (s_ser_rel, s_ser_find, s_ser_imp, s_ser_rec, m_ser_find, m_ser_imp, m_ser_rec) broaden user preferences more than the corresponding non-serendipitous ones.

To compare different kinds of serendipity with each other, we conducted direct comparisons between them. Similarly, to compare

Table 5: The fixed effects (coefficients) of the sixteen cumulative link mixed-effect regression models. Each cell corresponds to a coefficient of an ordinal regression with a single independent variable. Dependent variables are the metrics (preference broadening or satisfaction), while the independent variables correspond to whether a movie is serendipitous according to a particular definition (metric ~ serendipity). Significance codes: "*" < 0.0005

	s_ser_rel	s_ser_find	s_ser_imp	s_ser_rec	m_ser_rel	m_ser_find	m_ser_imp	m_ser_rec
Broadening	0.979*	1.471*	1.581*	1.605*	0.663	1.667*	1.354*	1.307*
Satisfaction	-0.347	0.322	0.284	0.276	-0.166	0.486	0.164	0.265

Table 6: The fixed effects (coefficients) of the twenty-eight cumulative link mixed-effect regression models. Each cell corresponds to a coefficient of an ordinal regression with the independent variable run on a dataset consisting of observations belonging to the kinds of serendipity indicated in the left column and the top row. The dependent variable corresponds to user satisfaction, while independent variables correspond to the variations of unexpectedness indicated in the left column (satisfaction ~ serendipity). Significance codes: "*" < 0.0005

Serendipity	s_ser_rel	s_ser_find	s_ser_imp	s_ser_rec	m_ser_rel	m_ser_find	m_ser_imp
s_ser_find	0.570						
s_ser_imp	0.503	-0.045					
s_ser_rec	0.071	0.092	-0.052				
m_ser_rel	0.185	-0.536	-0.445*	-0.520			
m_ser_find	0.733	0.047	0.140*	0.063	0.522		
m_ser_imp	0.514	-0.244	-0.091	0.017	0.274	-0.309	
m_ser_rec	1.069	-0.135	-0.145	0.355*	0.600	0.005	0.153

variations of serendipity components, we ran each comparison on an altered dataset, which included only observations of the two kinds of serendipity that were being compared and repeated observations belonging to both kinds simultaneously. Overall, we ran fifty-six cumulative link mixed-effect regression models (twenty-eight models per metric). We omitted the results for preference broadening, as we did not find any statistically significant results. Table 6 demonstrates the fixed effects (coefficients) of the regression models run for user satisfaction. The following observations can be noticed:

- m_ser_find movies are more enjoyable than s_ser_imp movies.
- s_ser_imp movies are more enjoyable than m_ser_rel movies.
- s_ser_rec movies are more enjoyable than m_ser_rec movies.

Motivational serendipity (find) outperforms strict serendipity (implicit), which, in turn, outperforms motivational serendipity (relevant) in terms of user satisfaction (observations 1 and 2). Meanwhile, strict serendipity (recommend) outperforms motivational serendipity (recommend) (observation 3).

5.3 Detecting Serendipitous Items

To answer RQ3, we come up with different features of movies and selected the effective subset for detecting serendipitous movies corresponding to the union of the six kinds of serendipity definitions, i.e. all except the two definitions s_ser_rel and m_ser_rel. We excluded these two from the union because (a) these two definitions include unexp_rel, for which we have evidence showing that it hurts user satisfaction and (b) we do not have evidence showing that m_ser_find broadens user preferences more than non m_ser_find. We predicted the union of the six serendipity definitions that broaden user preferences more than their corresponding non-serendipitous items.

To support feature calculation, we first define an average similarity of a movie to a user profile or to the recommendations this user previously received. In this paper, the term *user profile* refers to ratings this user assigned to items in the past. We define the average similarity of a movie to a user profile as follows:

$$sim\_prof_{u,i} = \frac{1}{||I_u||} \sum_{j \in I_u, j \neq i} sim_{i,j} \tag{2}$$

where I_u is the set of movies rated by user u, while $sim_{i,j}$ is the similarity between movies i and j. The way we calculate similarity depends on the movie representation. For example, to calculate genre similarity, we modeled movies as sets of genres and used the Jaccard similarity. For collaborative similarity, we modeled movies as rating vectors, where each value corresponded to a user rating, and cosine similarity was used. We define the average similarity of a movie to the recommendations a user previously received as follows:

$$sim\_rec_{u,i} = \frac{1}{||R_u||} \sum_{j \in R_u, j \neq i} sim_{i,j} \tag{3}$$

where R_u is the set of the eight last movies recommended to user u by MovieLens. In summary, we came up with the following features:

- Popularity (logpop). We used popularity because it is one of the most common attributes used in studies dedicated to serendipity in recommender systems [19, 20, 30]. We calculated popularity as follows: $logpop_i = ln(U_i)$, where U_i is the number of ratings received by movie i during the last year (2016) in MovieLens. We picked the number of ratings during the last year instead of the overall number of ratings, because many old movies received many ratings if they were released a long time ago. However, these movies were likely to be unfamiliar to the active users in the system. The most famous movies, such as "The Shawshank Redemption", "Toy Story" and "The Matrix" are still among the most popular movies according to our last-year popularity metric.

- Predicted rating (predicted_rating). We used this feature because the expectation of users might be affected by the system's predictions, while they are browsing movie pages (note that in MovieLens, the predicted ratings are displayed along with the movie information). The algorithm that predicts the rating depends on the choice of the user because MovieLens offers several recommendation algorithms, among which item-based collaborative filtering and matrix factorization are used by the majority of the users.
- Release year (year). We picked this attribute because recency of movies might affect users' familiarity with them. Users might be more familiar with recently released movies than the older ones.
- Average tag-based similarity to the user profile (tag_sim_prof). Similarly to popularity, we picked this feature because content-based similarity is commonly considered in the literature [15, 19, 29, 30]. To calculate the average tag-based distance we employed the tagging model, tag genome [28], which is based on tags users assign to movies. We calculated the distance according to Equation 2, where $sim_{i,j}$ is the similarity measure of weighted cosine distance in [28].
- Average tag-based similarity to recommendations the user received from MovieLens (tag_sim_rec). We picked this feature because users' expectation might depend on the recommendations our system generates. We calculated this feature according to Equation 3 using the similarity measure of weighted cosine distance in [28].
- Average genre-based similarity to the user profile (genre_sim_prof). We picked this feature as an additional content-based similarity and calculated it according to Equation 2, where $sim_{i,j}$ is the Jaccard similarity between the sets of genres of the movies i and j.
- Average genre-based similarity to recommendations the user received from the system (genre_sim_rec). We picked this feature as an additional content-based similarity and calculated it according to Equation 3 using the Jaccard similarity.
- Average collaborative similarity to user profile (c_sim_prof). We picked this feature because this is a common similarity measure in the literature on serendipity [14, 30]. We calculated this feature according to Equation 2, where $sim_{i,j}$ is the cosine similarity between movie rating vectors i and j.
- Average collaborative similarity to recommendations the user received from the system (c_sim_rec). We calculated this feature according to Equation 3 using the cosine similarity.

We detected effective features for predicting serendipitous movies by running a logistic regression model on our dataset. We used the logistic regression model for the sake of interpretability. In our dataset, we labeled each movie based on whether this movie was serendipitous to a user and performed the 10-fold cross validation.

To select effective features for the prediction of serendipity, we employed the forward search strategy, where we iteratively picked features based on the performance of the logistic regression model when gradually adding these features into the model. To compare models, we used the metric: Area Under the ROC Curve (AUC), which is a commonly used for assessing performance of binary classifiers. We also reported AIC (Akaike Information Criterion), which evaluates the quality of a statistical model (the lower the value, the better the model) [3].

Table 7: The results of feature selection with logistic regression, where the dependent variable is a binary variable indicating whether a movie belongs to the union of the six kinds of serendipity (excluding s_ser_rel and m_ser_rel).

Features	AIC	AUC
predicted_rating	1468.628	0.609
predicted_rating + logpop	1464.008	0.621
predicted_rating + logpop + tag_sim_prof	1459.707	0.624
predicted_rating + logpop + tag_sim_prof + c_sim_prof	1459.840	0.627

Table 7 demonstrates the results of the forward feature selection strategy. We only included the first four features, because further incorporating more features decreases AUC. According to the obtained results, the most effective features for serendipitous movies according to at least one of the six definitions are predicted rating, popularity, the average tag-based similarity to the user profile and the average collaborative similarity to the user profile.

Table 8: The coefficients of the logistic regression model, where the dependent variable is a binary variable indicating whether a movie is serendipitous according to the union of the six definitions, while the four independent variables correspond to the selected features

Feature	Parameter	Standard Error
predicted_rating	2.954*	0.624
logpop	-1.223	0.392
tag_sim_prof	0.508	0.260
c_sim_prof	-0.816	1.057

Table 8 shows the coefficients of the final logistic regression model. It shows that movies that are serendipitous according to at least one of the six definitions have higher predicted ratings than corresponding non-serendipitous movies. Other coefficients are not statistically significant after correction, but the model shows a trend that serendipitous movies tend to be less popular compared with non-serendipitous ones.

5.4 How Rare Are Serendipitous Items?

According to Table 2, among 2146 movies that users gave their feedback on, 302 (14%) are serendipitous according to at least one definition. The entire database of MovieLens contains 25,650,696 ratings and 15,854,339 (or 61%) of them are higher than 3.5, which suggests that up to 8.5% ($0.14 * 0.61 \approx 0.085$) are serendipitous. Our samples include 437 movies that our system encouraged users to watch, which can be considered as the number of recommendations that users took. This suggests that up to 69% of recommendations provided by our system that users watch are serendipitous according to at least one definition.

The dataset includes 275 movies that correspond to the union of the six definitions of serendipity, which have a positive effect on preference broadening. This suggests that our system contains up to 5.9% of these movies and 47.4% of them among the recommendations. For the smallest kind of serendipity, strict serendipity (recommend), these ratios are 1.8% and 14.4%, while for the largest kind of serendipity, strict serendipity (find), they are 5.1% and 41.4%, respectively.

6 DISCUSSION

We reviewed a set of techniques for operationalizing serendipity, finding that different definitions have different effects on preference broadening and user satisfaction, but confirming that in general serendipitous recommendations broaden preferences (usually without hurting satisfaction).

We found that there are sufficient serendipitous items to recommend (particularly across the span of definitions), making it feasible to recommend serendipitous items in contexts where preference broadening would be useful. We do not look explicitly at which contexts may benefit most, but leave that to others.

The results of our study regarding features effective for the detection of serendipitous items mostly correspond to the prior literature. Content-based similarity to a user profile, collaborative similarity to a user profile and popularity have been acknowledged as important features [1, 14, 15, 20, 22, 29, 30]. However, most studies employ popularity and disregard similarity to a user profile in offline evaluations of recommendation algorithms [20, 22, 30].

Surprisingly, our results showed that ratings provided by recommender systems are a good predictor for serendipity. In fact, the higher the rating the more likely an item to be perceived serendipitous. This might suggest that even recommendation algorithms optimized for accuracy assist users to encounter serendipitous items at least within the limitations of our dataset. This contradicts to the common claim that recommender systems narrow users' interests and trap them in filter bubbles [23, 27, 29]. However, the design of our experiment does not allow us to support or reject this claim.

7 LIMITATIONS AND FUTURE WORK

Conducting a study of movie recommender system users based on their previously-rated movies has several limitations. First, we were limited in the reasons we could explore for movie performance on our metrics. For example, serendipitous movies might broaden user preferences more than non-serendipitous ones due to other reasons than that these movies are serendipitous. Second, our study is limited in domain to movies. While we hope our results are generalizable at least in related domains, further study is needed to evaluate user impact, even in this domain. In our future work, we are going to design a serendipity-oriented algorithm using the collected dataset and evaluate it in an experimental setting with real users, where we control for serendipity with the novel algorithm.

The specific design of our study had other limitations. We only looked at performance of different kinds of items in terms of preference broadening and user satisfaction, which was based on the literature review. Future work should consider other metrics. We selected only relatively unpopular relevant movies for our survey to increase the chance of asking users about serendipitous movies, which only allowed us to compare unpopular serendipitous movies and unpopular relevant non-serendipitous ones. As a result, our sample is biased, and may not represent average performance. Finally we limited our study to active users (duration of use of at least a month, minimum number of ratings), which may not reflect the experience of one-time or very infrequent users.

8 CONCLUSION

In this paper, we conducted a survey asking 475 real users about 2146 movies with questions designed based on different serendipity definitions synthesized from the prior literature. Through this survey, we collected the first dataset that has real user evaluation on the serendipity of the items. We only asked about relevant movies to (i.e. highly rated by) those users and therefore the effects we found in this work are all relative to items that are relevant but not serendipitous. The following research questions are addressed.

RQ1. What are the effects of various serendipity components on broadening user preferences and user satisfaction?

We found that all variations of the unexpectedness and novelty components broaden user preferences, but one type of unexpectedness (unexpected to be relevant) hurts user satisfaction. Movies that users found novel and unexpected according to any definition employed in this paper broaden user preferences more than movies users found non-novel and non-unexpected, respectively. Movies that users did not expect to like and be recommended are less enjoyable than movies users expected to like and be recommended (or had no expectations), respectively.

Variations of the unexpectedness component are different in terms of our metrics. Two variations of unexpected movies: (a) movies that users did not expect to find and (b) movies that users thought were different from movies these users usually watch are better than the variation: (c) movies users did not expect to like in terms of both preference broadening and user satisfaction. Meanwhile, movies that users did not expect to find are more enjoyable than the ones that users found different from movies these users usually watch.

RQ2. What are the effects of different kinds of serendipity on broadening user preferences and user satisfaction?

We found that serendipitous movies generally broaden user preferences more than non-serendipitous ones, but we did not find any effects of serendipity on user satisfaction. In particular, movies that are serendipitous according to seven definitions of serendipity broaden user preferences more than corresponding non-serendipitous ones.

We also found that different kinds of serendipity differ in terms of user satisfaction. In particular, motivational serendipity (find) outperforms strict serendipity (implicit), which, in turn, outperforms motivational serendipity (relevant), while strict serendipity (recommend) outperforms motivational serendipity (recommend).

RQ3. What are the effective features for detecting serendipitous movies? What are the value ranges of these features for typical serendipitous movies?

We found features most important for detecting movies that are serendipitous according to six definitions of serendipity that do not include unexpectedness (relevant), which hurts user satisfaction. These features are predicted ratings, popularity, content-based and collaborative similarity to a user profile. Our results also show that these serendipitous movies have higher predicted ratings than corresponding non-serendipitous ones.

RQ4. How rare are serendipitous movies among movies rated by the users in a typical collaborative-filtering-based recommender system? To what extent does this kind of system help users find these movies?

We discovered that in the best case scenario, among movies users rate in a typical movie recommender system, up to 8.5% are serendipitous according to at least one definition, while among movies recommended by the system that users watch, this ratio is up to 69%. We only provide an upper bound estimation due to the bias of our dataset.

ACKNOWLEDGMENTS

The research at the University of Jyväskylä was partially supported by the Academy of Finland, grant #268078 and the KAUTE Foundation.

We would like to thank F. Maxwell Harper and researchers at the GroupLens team and the MineSocMed project for their feedback and assistance in conducting this research.

REFERENCES

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. 2014. On Unexpectedness in Recommender Systems: Or How to Better Expect the Unexpected. *ACM Transactions on Intelligent Systems and Technology* 5, 4 (2014), 1–32.

[2] Alan Agresti and Maria Kateri. 2011. Categorical data analysis. In *International encyclopedia of statistical science*. Springer, 206–208.

[3] Hirotugu Akaike. 1974. A new look at the statistical model identification. *IEEE transactions on automatic control* 19, 6 (1974), 716–723.

[4] Paul André, M.C. Schraefel, Jaime Teevan, and Susan T. Dumais. 2009. Discovery is Never by Chance: Designing for (Un)Serendipity. In *Proceedings of the Seventh ACM Conference on Creativity and Cognition*. ACM, New York, NY, USA, 305–314.

[5] Rune Haubo Bojesen Christensen. 2010. ordinal-regression models for ordinal data. *R package version* 22 (2010).

[6] Marco de Gemmis, Pasquale Lops, Giovanni Semeraro, and Cataldo Musto. 2015. An investigation on the serendipity problem in recommender systems. *Information Processing & Management* 51, 5 (2015), 695 – 717.

[7] Michael D Ekstrand, Daniel Kluver, F Maxwell Harper, and Joseph A Konstan. 2015. Letting users choose recommender algorithms: An experimental study. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 11–18.

[8] Michael D. Ekstrand, John T. Riedl, and Joseph A. Konstan. 2011. Collaborative Filtering Recommender Systems. *Foundations and Trends in Human-Computer Interaction* 4, 2 (2011), 81–173.

[9] Mouzhi Ge, Carla Delgado-Battenfeld, and Dietmar Jannach. 2010. Beyond Accuracy: Evaluating Recommender Systems by Coverage and Serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*. ACM, 257–260.

[10] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, and John T. Riedl. 2004. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22, 1 (2004), 5–53.

[11] Yoshinori Hijikata, Takuya Shimizu, and Shogo Nishida. 2009. Discovery-oriented Collaborative Filtering for Improving User Satisfaction. In *Proceedings of the 14th International Conference on Intelligent User Interfaces*. ACM, 67–76.

[12] Yosef Hochberg and Yoav Benjamini. 1990. More powerful procedures for multiple significance testing. *Statistics in medicine* 9, 7 (1990), 811–818.

[13] Leo Iaquinta, Giovanni Semeraro, Marco de Gemmis, Pasquale Lops, and Piero Molino. 2010. *Can a recommender system induce serendipitous encounters?* InTech.

[14] Marius Kaminskas and Derek Bridge. 2014. Measuring Surprise in Recommender Systems. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design (Workshop Programme of the 8th ACM Conference on Recommender Systems)*.

[15] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Transactions on. Interactive Intelligent Systems* 7, 1, Article 2 (2016), 42 pages.

[16] Komal Kapoor, Vikas Kumar, Loren Terveen, Joseph A. Konstan, and Paul Schrater. 2015. "I Like to Explore Sometimes": Adapting to Dynamic User Novelty Preferences. In *Proceedings of the 9th ACM Conference on Recommender Systems*. ACM, 19–26.

[17] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2016. Challenges of Serendipity in Recommender Systems. In *Proceedings of the 12th International conference on web information systems and technologies.*, Vol. 2. SCITEPRESS, 251–256.

[18] Denis Kotkov, Jari Veijalainen, and Shuaiqiang Wang. 2017. A Serendipity-Oriented Greedy Algorithm for Recommendations. In *Proceedings of the 13th International Conference on Web Information Systems and Technologies*, Vol. 1. ScitePress, 32–40.

[19] Denis Kotkov, Shuaiqiang Wang, and Jari Veijalainen. 2016. A survey of serendipity in recommender systems. *Knowledge-Based Systems* 111 (2016), 180–192.

[20] Qiuxia Lu, Tianqi Chen, Weinan Zhang, Diyi Yang, and Yong Yu. 2012. Serendipitous Personalized Ranking for Top-N Recommendation. In *Proceedings of the The IEEE/WIC/ACM International Joint Conferences on Web Intelligence and Intelligent Agent Technology*, Vol. 1. IEEE Computer Society, 258–265.

[21] Sean M. McNee, John Riedl, and Joseph A. Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*. ACM, 1097–1101.

[22] Tomoko Murakami, Koichiro Mori, and Ryohei Orihara. 2008. Metrics for Evaluating the Serendipity of Recommendation Lists. In *Annual Conference of the Japanese Society for Artificial Intelligence*.

[23] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

[24] Paul Resnick and Hal R. Varian. 1997. Recommender Systems. *Commun. ACM* 40, 3 (1997), 56–58.

[25] Alan Said, Ben Fields, Brijnesh J. Jain, and Sahin Albayrak. 2013. User-centric Evaluation of a K-furthest Neighbor Collaborative Filtering Recommender Algorithm. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work*. ACM, 1399–1408.

[26] Eugenio Tacchini. 2012. *Serendipitous mentorship in music recommender systems*. Ph.D. Dissertation.

[27] Maria Taramigkou, Efthimios Bothos, Konstantinos Christidis, Dimitris Apostolou, and Gregoris Mentzas. 2013. Escape the Bubble: Guided Exploration of Music Preferences for Serendipity and Novelty. In *Proceedings of the 7th ACM Conference on Recommender Systems*. ACM, 335–338.

[28] Jesse Vig, Shilad Sen, and John Riedl. 2012. The Tag Genome: Encoding Community Knowledge to Support Novel Interaction. *ACM Transactions on Interactive Intelligent Systems* 2, 3, Article 13 (2012), 44 pages.

[29] Yuan Cao Zhang, Diarmuid Ó Séaghdha, Daniele Quercia, and Tamas Jambor. 2012. Auralist: Introducing Serendipity into Music Recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*. ACM, 13–22.

[30] Qianru Zheng, Chi-Kong Chan, and Horace H.S. Ip. 2015. An Unexpectedness-Augmented Utility Model for Making Serendipitous Recommendation. In *Advances in Data Mining: Applications and Theoretical Aspects*. Vol. 9165. Springer International Publishing, 216–230.