SIRUP: Serendipity In Recommendations via User Perceptions

Valentina Maccatrozzo
Vrije Universiteit Amsterdam
The Netherlands
v.maccatrozzo@vu.nl

Manon Terstall
Vrije Universiteit Amsterdam
The Netherlands
m.terstall@student.vu.nl

Lora Aroyo
Vrije Universiteit Amsterdam
The Netherlands
lora.aroyo@vu.nl

Guus Schreiber
Vrije Universiteit Amsterdam
The Netherlands
guus.schreiber@vu.nl

ABSTRACT

In this paper, we propose a model to operationalise serendipity in content-based recommender systems. The model, called SIRUP, is inspired by the Silvia's curiosity theory, based on the fundamental theory of Berlyne, aims at (1) measuring the novelty of an item with respect to the user profile, and (2) assessing whether the user is able to manage such level of novelty (coping potential). The novelty of items is calculated with cosine similarities between items, using Linked Open Data paths. The coping potential of users is estimated by measuring the diversity of the items in the user profile. We deployed and evaluated the SIRUP model in a use case with TV recommender using BBC programs dataset. Results show that the SIRUP model allows us to identify serendipitous recommendations, and, at the same time, to have 71% precision.

ACM Classification Keywords

H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous; H.1.2 User/Machine Systems: Human information processing

Author Keywords

Personalization;Television/Video; Entertainment; Design Methods; Qualitative Methods; User and Cognitive models; User Studies; Serendipity; Recommender System; Curiosity

INTRODUCTION

TV broadcasters are more and more providing their content on the Web via new (mobile devices) apps [14]. Also, third party companies, such as Amazon Prime Video[1], Netflix[2], and many others, are providing online TV content. This gives the possibility to the users to watch TV at any time, from any place, enhancing engagement with content. In the Netherlands, for example, Internet users watching TV and listening to the radio online were 63% last year [4]. In the U.S. 78% of smartphones users access TV apps at least once last January [13].

To manage such an abundance of content, there is a high need of recommender systems to filter out the content for users. However, recommender systems work well only when information at their disposal, both about users and content, is up-to-date and as thorough as possible [6]. When this does not happen, we get into cold start [29] and filter bubble problems [27].

The filter bubble problem is caused by a recommender system when it displays to the user only content very similar to her profile. This prevents the user from discovering different and new content, and to expand her horizon of interests. Relaxing the filters is not enough to address the problem, as this approach does not take into account explicitly for the fact that recommendations need to be novel and diverse from the user profile, but still relevant to it. In this paper, we propose to address the filter bubble problem by introducing a model for serendipity in a content-based recommender system.

Serendipity is a complex concept when it comes to being explained in words. The original definition is "*[...] making discoveries, by accidents and sagacity, of things which they were not in quest for [...]*" [34]. We started from this definition to define our working definition of serendipity: "making a pleasant and relevant discovery that was unexpected". **Pleasant** as it has to have a positive connotation; **relevant** as it has to be linked to the person's knowledge; **unexpected** to account for the surprise aspect. We apply this working definition in the context of TV recommender systems, making the assumption that, when it comes to TV programmes, a recommender system is used with explorative purposes: if the user knows what she wants to watch, she does not need to use a recommender system.

A serendipitous recommendation brings the user to an unknown item that she would most likely not have discovered autonomously, but is perceived by her as being surprisingly

[1] https://www.amazon.com/Prime-Video/b?node=2676882011
[2] https://www.netflix.com

IUI 2017, March 13 - 16, 2017, Limassol, Cyprus

ISBN 978-1-4503-4348-0/17/03...$15.00

DOI: http://dx.doi.org/10.1145/3025171.3025185

interesting. In other words, the serendipitous recommendation triggers a positive effective state in the user (interest) that motivates her to follow the recommendation (*e.g.*, she wants to know more about the recommended item). Kashdan and Silvia [18, 32] describe this state as curiosity: "the recognition, pursuit, and intense desire to explore novel, challenging, and uncertain events". The recommended TV programme would be the novel, challenging, and uncertain event for which desire of exploration is induced by the recommender system. This brings us to our first research question:

RQ1: *Do serendipitous recommendations trigger curiosity in users?*

In this paper, we introduce a serendipity model (SIRUP) based on psychological curiosity theories. We base SIRUP on the theory of Berlyne [2], which has been more recently renewed and extended by Silvia. In particular, according to Silvia the emotion of interest (which he defines the "curious emotion" [32]) had two appraisals: (1) the appraisal of something as new, unexpected, or complex, and (2) an appraisal of one's ability to comprehend the new, complex thing [31]. We define the first appraisal the *novelty check* and the latter appraisal the *coping potential check*. These two components constitute the main elements of SIRUP. It is important to notice that the novelty check focuses on the data, by measuring the novelty of a TV programme with respect to the items in the user profile, while the coping potential check focuses on the user's ability to manage such level of novelty.

It has been proven that Linked Open Data (LOD) is a good data source for serendipity in exploratory search [11, 15, 33]. In particular, through traversing data links, systems may, not only use up-to-date data but also automatically discover unknown content [15]. Starting from these claims, we propose to use LOD paths[3] to surface new, serendipitous connections between TV programmes, and use these paths as input in SIRUP. This translates into our second research question:

RQ2: *Can we perform the novelty check of TV programmes with respect to the user profile using LOD paths components?*

We use semantic enrichment [10, 16, 21, 22, 25, 26] as a means to link TV programmes with LOD resources.

In order to perform the coping potential check, we need to make an estimation based on what we know about the users. The coping potential is an ability of the user of dealing with different topics, so we propose to estimate it by measuring the diversity of the TV programmes in the user profile. This allows us to state our third research question:

RQ3: *Can we estimate the coping potential of a user with the diversity of the TV programmes in the user profile?*

To test the SIRUP model, we instrumented an experiment using a dataset of BBC programmes. Results show that the SIRUP model successfully allows us to identify serendipitous recommendations, without harming too much precision, which reaches 71%.

The contribution of this paper is twofold: (1) a model to define serendipity in recommender systems, and (2) the operationalisation of such model using semantic enrichment and LOD paths.

The paper continues with related work, followed by the introduction and definition of the SIRUP model. After describing the experiment, we present the results and discuss them. In the last section, we conclude and give an indication for future directions.

RELATED WORK

The main idea behind the Semantic Web is to allow reuse of ontologies and data, which is constantly enriched and improved. It allows the users to engage in a serendipitous experience by discovering related information [30]. This claim has been taken into account especially for exploratory search systems. A first attempt has been presented by Dimitrova *et al.* [11], who investigate the usage of semantic tags and their effect in an exploratory search browser. They instrumented an experiment using data from DBpedia, DBTune and Amazon reviews with the aim to facilitate exploratory search about music instruments. Their results indicate that the semantic tags support the task of exploratory search and facilitate serendipitous learning. The underlying idea that drove this research is the same as ours, however, this work focusses on serendipitous learning via user exploration, while we focus on suggesting serendipitous TV programmes without the user aid. Another example of LOD to support exploratory search is presented by Marie *et al.* [23]. They explore the LOD graph by generating paths with spreading activation combined with a sampling technique to be able to compute results *on-the-fly*. Their algorithm features a *Serendipity Mode*, which retrieves results adding randomness. They performed a user experiment to test several hypotheses, including the evaluation of the ranked list of resources for relevance and discovery score. Their algorithm performed well in both scores. To allow the process *on-the-fly* they had to sample the graph, which we prefer to avoid in order to guarantee the inclusion of all possible paths, to allow the system to identify the more serendipitous connections.

One application in the bio-medical domain, presented by Saleem *et al.* [28] shows the potential of LOD for serendipitous hypothesis making. They link medical LOD dataset to allow for the surface of not-so-easy-to-see connections between patients and treatments. However, they test their system for feasibility and not for actual serendipitous hypothesis making. Nevertheless, this is an interesting approach which shares the idea of surfacing serendipitous connections between LOD resources.

Two works which propose LOD paths-based techniques for serendipity are the ones of Dojchinovski and Vitvar [12] and of De Vocht *et al.* [9]. Dojchinovski and Vitvar propose a collaborative-based personalisation method to suggest resources of interest to users. They use a path-based similarity calculation between users by generating resources context graph, with distance equal to two. They perform an offline

[3]LOD paths are characterised by the types and relationships of the objects which compose them.

evaluation of this algorithm compared to other state of the art algorithms and show it improves accuracy, serendipity and diversity. We identify several differences with our approach. First, we also use paths but we take into account also properties included in them; second, we use paths with up to length three, and, finally, we define serendipity as a qualitative measure, and as such, we evaluate it with a user experiment. The work of De Vocht *et al.* focuses on LOD path-based story telling. They use several heuristics and weights to recalibrate a previously presented algorithm in order to guarantee more coherent paths. They find that some heuristics and weighing schema do increase discovery, but no consistency is found in terms of relevance. As these results show, this approach is too objective and, as such, misses the subjectivity of serendipity.

One of the first work on serendipity in recommender systems is the one from Iaquinta *et al.* [17]. They identify serendipitous documents as the ones for which the system is more uncertain, *i.e.*, the item for which the system is not able to generate a rating. Their results show that by increasing randomness in choosing items from the serendipitous list the users' ratings increases as well. However, they are not sure whether the topic of the suggested documents was completely unknown to the users. Our approach differentiates in the fact that we do not make use of any randomness in the selection of the serendipitous items, rather we carefully select it based on the user ability to deal with new content. A nice analysis of serendipitous items has been presented by Akiyama *et al.* [1]. They collected data from users by asking them to indicate for several TV programmes whether it was 'interesting and recognised', 'not interesting' or 'serendipitous'. In the gathered data they found that non-interesting items were distributed outside the recognised area of interest and serendipitous recommendations are distributed far outside this area. Based on this analysis, they propose a distance metric to make serendipitous recommendations. In this work, while we recognise the existence of the recognised area of interest, we propose a method to actually understand how big this area is for every user and how much outside of it we need to go to find serendipitous items. Another interesting work about serendipity is the one from Zhang *et al.* [35], who introduce the Auralist recommendation framework. This framework tries to balance and improve accuracy, diversity, novelty and serendipity simultaneously. They try to identify "user's local preference graph" and identify potentially serendipitous items outside of it. This approach successfully enhances serendipity, novelty and diversity with little harm on accuracy. One of their main conclusions is that serendipity increases user satisfaction. This approach recognises the existence of the user bubble too, but they also do not try to estimate how much outside of it they need to go to find serendipitous items, as we try to do. A more recent work about serendipity in recommender system is the one from de Gemmis *et al.* [8]. The authors propose a strategy that enriches a graph-based recommendation algorithm with background knowledge (WordNet and Wikipedia). They show that this additional knowledge actually allows them to introduce non-obvious recommendations to users, without harnessing too much accuracy. This approach shares some similarities with our approach, but it does not make use of structured knowledge, as we do, and, it does not personalise the serendipity approach, *i.e.*, they identify correlations between keywords to identify serendipitous recommendations, without taking into consideration the user interest area. Another interesting work which does not focus on serendipity but it is relevant to our work is the one from Cremonesi *et al.* [5]. The authors perform a deep comparison of 7 recommender system algorithms and prove that objective measures, like F-measure, do not always grasp the quality perceived by the user. Besides, they propose a new definition of novelty: first and second order novelty. The first is novelty, in terms of movies, deals with the fact that the user did not watch the movie, while the latter is a more strict definition of novelty as a movie completely unknown to the user (never heard of it). This second definition is stated as a definition of serendipity. In our work, we define serendipity more as the first order novelty, however we argue that novelty is only one aspect of serendipity. Our working definition indeed accounts also for relevance and the unexpectedness of serendipity.

SIRUP

The subjectivity of serendipity depends mainly on two factors: the knowledge of the user and how much the user is keen on knowing more. The latter is what we are used to calling curiosity: a strong desire to know or learn something[4].

Indeed, the user will not get curious about something she already knows, *i.e.*, recommend an item she is familiar with; and the user will not get curious about something she does not know at all, *i.e.*, a recommendation irrelevant to her interest area. Psychological theories describe curiosity as an internal conflict [2], *i.e.*, a gap which arises when there is a discrepancy between the current knowledge level and the desired knowledge state. Only when the gap has the right size, *i.e.*, it is manageable, the curiosity of the person will be triggered [19]. Silvia describes interest (which he defines as the 'curious emotion') in terms of appraisals: (1) a novelty-complexity check and a (2) coping-potential check [32]. Interest is induced by new, complex, and unfamiliar events [3] when people feel able to deal with the challenges that they pose [7].

We can map these two aspects to our working definition of serendipity, *i.e.*, "making a pleasant and relevant discovery that was unexpected": unexpected maps back to the new and unfamiliar event, while pleasant and relevant map to the trust users have in themselves to deal with the unexpected event. In other words, if the user feels like she is able to deal with the unexpected event, then the event is pleasant to her. SIRUP comprises both these aspects by implementing two elements, (see Figure 1):

1. a novelty potential check;
2. a coping potential check.

The first assesses the item's novelty with respect to the items in the user profile. This check focuses on the characteristics of the data, it is an objective measurement. The latter focuses on assessing the ability of the user to deal with such amount of novelty. This is a qualitative measurement: we can partly

[4] https://en.oxforddictionaries.com/definition/curiosity

deduce it from what we know about the user, but it is mainly a personal attitude.

Novelty check

The novelty check aims at assessing the novelty of an item to recommend with respect to the items in the user profile. It is very similar to the core of a standard recommender system, but it ranks items based on the novelty: *i.e.*, the more different from the user profile, the higher in the rank. Novelty can be seen as the inverse of similarity: if it is less similar to the items in the user profile, it is perceived as more novel.

Given the previous successful uses of LOD for serendipity [11, 12], we propose to use LOD paths to measure the novelty of the items to recommend. In particular, we focus on the structure of these paths, *i.e.*, types and properties which compose them. We use the cosine similarity measure (see Equation 1) for this purpose. The cosine similarity measure has several advantages, including the fact that it allows us to compare also vectors of different lengths, which happens in our case as we are using LOD paths of different lengths.

$$cosine\_similarity = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n}(A_i^2)} \times \sqrt{\sum_{i=1}^{n}(B_i^2)}} \quad (1)$$

We introduced the usage of LOD paths in content-based recommender system in a previous work [20]. An LOD path is an ordered set of types and properties, which connects two types, T_1 and T_{l+1}:

$$\{T_1, P_1, T_2, P_2, ..., T_l, P_l, T_{l+1}\}$$

where l is the length of the pattern. In order to be able to extract these patterns, we first need a link between our items and an LOD dataset. In our case, we perform semantic enrichment of the title of the TV programme with DBpedia[5] concepts. Then, we extract patterns from the DBpedia knowledge space, between the aligned concepts, up to length of 3.

LOD paths bring several advantages. First, they allow us to discover connections between items that otherwise would not have been discovered [20]. For example, LOD paths allow us to link the documentary *Reggie Yates's Extreme South Africa*[6] to the show *The Sky at Night*[7]. The link is found through the word *extreme*, which is associated with the musical band Extreme. This band is influenced by the band 'Queen' whose member was Brian May, who was a guest in the show *The Sky at Night*.

The similarity between a TV programme to recommend and the TV programmes in the user profile using LOD paths components is calculated as follows. When there exist LOD paths which connect the TV programme to recommend and one of the TV programme in the user profile, we use the types and properties that constitute these paths as input for the cosine similarity, like they are keywords that describe the TV programme. If there are no LOD paths that connects the two TV programmes, their similarity is zero. The rationale behind this choice is that we expect that users unconsciously follow specific behavioural patterns when choosing a TV programme to watch, *e.g.*, their favourite actor is in the TV programme or the TV programme is of their favourite genre, and so on. Through LOD paths we are able to identify all possible patterns, and, our idea, is that the more diverse this recurrent patterns that connect two TV programmes are, the more the recommended TV programme will appear as novel to the user.

Coping Potential Check

Estimating the coping potential of the user from her profile is challenging. First of all, we always have an incomplete knowledge of the user preferences: even if we know what the user usually watches, we do not have access to other interests of the user which could potentially affect her watching behaviour. Additionally, preferences changes over time, so profiles need to be updated, Second, it is difficult to estimate the user attitude towards new content, if, for instance, the user watches always and only the same TV programme (*e.g.*, different episodes of a TV series). We believe that a proper assessment of the user's coping potential should be performed through a questionnaire. In this paper, we propose a simplified approach, by estimating the coping potential of the users with the diversity liked genres and formats within the user profile. In particular, we propose to count the unique instances of genres and formats and use this number as an indicator of the coping potential. Our aim is to classify users in two categories: low and high coping potential. Our idea is that the more diversity there is in what the user likes, the more she is able to cope with different content and, so, has a wider range of interests and a higher coping potential.

EXPERIMENT

We performed an online experiment to test the SIRUP model. We set up a job on the crowdsourcing platform CrowdFlower[8]. The job was launched on 28^{th} October 2015.

Goal

The experiment has been designed in order to answer the following research questions:

1. Can familiarity be measured using the cosine similarity based on LOD paths components?
2. Can the user coping potential be estimated by calculating the diversity of genres and formats within the user profile?

Participants

In total 290 people took part to the experiment: 202 participants completed the questionnaire, the remaining 14 participants rated on average 6 recommendations. Besides, based on a control question, we excluded 51 participants (see Section Survey structure for more details about the control question). The analysis presented in this paper are based on 165 participants, and all of them completed the questionnaire. More details about the participants can be found in Table 1. We required only British participants, as we use BBC TV programmes for the experiment.

[5] http://dbpedia.org/
[6] http://www.bbc.co.uk/programmes/b03w79fx
[7] http://www.bbc.co.uk/programmes/b006mk7h
[8] http://www.crowdflower.com/

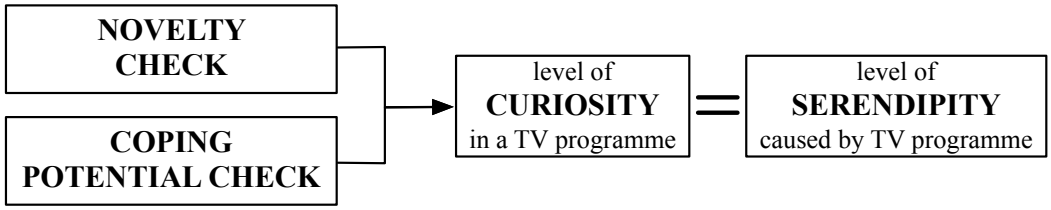

Figure 1. Serendipity Model

Age	Men	Women
< 20	4	0
20-29	10	25
30-39	18	21
40-49	25	17
50-59	15	14
≥ 60	6	4
Total	78	81

Table 1. Participants demographics information.

Data

We used BBC TV programmes aired between 7^{th} September 2015 and 20^{th} September 2015. We excluded Children's programmes and News bulletins. The first because we had only adults taking part to the study, the latter to avoid recommending TV programmes such as the weather forecast.

In total the dataset consists of 1460 TV programmes. Tables 3 and 2 show respectively the distribution of genre and format of the TV programmes.

Format	Total
Animation	55
Appeals	2
Discussion	49
Docudramas	1
Documentaries	259
Films	22
Games & Quizzes	109
Magazines & Reviews	83
Makeovers	14
Performance & events	40
Reality	24
Undefined	803
Total	1461

Table 2. Distribution format TV episodes

Survey structure

The experiment was set up as an online survey. We recruited participants through CrowdFlower. The survey was composed of three parts:

1. **Ratings.** Users are asked to rate 8 TV programmes. They also have the option to indicate whether they do not know them. In case none of the displayed TV programmes are known, another selection of TV programmes is shown, until at least one programmes is found to be known. To guarantee a good coverage of genres, we randomly selected one TV programmes per genre (the only genre we exclude is religion given its limited presence in the dataset).
2. **Favourite genre and format & demographics.** Users are asked to indicate their favourite genre(s) and format(s) and to answer few personal questions (*e.g.*, age, gender, and so on).
3. **Recommendations evaluation.** Users are asked to rate 18 recommendations along three dimensions (*e.g.*, interest, unexpectedness and relevance) with 5 points Likert scale questions. We mapped these three dimensions with the following statements:

   Question 1: I did not think of this TV programme, but it seems interesting to me. (Interest)

   Question 2 (control question): This TV programme does not seem interesting to me. (Interest)

   Question 3: I am surprised to get this TV programme recommended. (Unexpectedness)

   Question 4: This recommendation fits my personal preferences. (Relevance)

Genre	Total
Comedy	147
Drama	84
Entertainment	120
Factual	725
Music	29
News	280
Religion	3
Sport	45
Weather	28
Total	1461

Table 3. Distribution genre TV episodes

The control question is used to identify possible spammers. When users give the same answer to question 1 and 2, we know they did not read the statements, and gave random answers.

The first two parts of the survey are used to build the user profile, which is used to generate the recommendations as explained in the next section.

Recommendations generation

In order to generate personalised recommendations, we collected some information about the preferences of the users. Then, we used this information to rank the recommendations.

First, for each TV programme in the dataset we calculate the cosine similarity to the liked TV programmes. In case the TV programme has a genre or format which the user indicated she likes, we add a weight of 0.2 to the similarity value (0.4 in case both genre and format are liked). The weight is chosen to guarantee a similarity value higher than average. We use the same approach also with the disliked TV programmes in the user profile. The ranking of the recommendations based on the disliked programmes is used to avoid recommending TV programmes which are more similar to the disliked items than to the liked ones. When this happens, these TV programmes are not recommended.

For testing purposes, we generate three different rankings. The first one is used as a baseline, and it is calculated using only BBC metadata as input data to the cosine similarity measure. The second one is based on LOD paths components, while the third one is calculated using together BBC metadata and LOD paths components. We randomly select 2 TV programmes to recommend per interval, to ensure the collection of enough observations for low, medium and high similarity values. In total users rated 18 recommendations.

RESULTS

This section describes the analysis of the results obtained with the experiment described in the previous section. The results are presented divided by the three different methods we used to generate recommendations: Baseline (BBC metadata), SIRUP (LOD paths components), and Combined approach (BBC metadata together LOD paths components). For every analysis we apply the following rationale:

- To assess interest in the recommendations, we use the control question reversed (*i.e.*, if the answer was five in the Likert scale, it becomes one, and so on). Question 1 contains two statements which we cannot map to a unique answer. So we excluded the answers to this question from the analysis.

- To identify the serendipitous recommendations we require that all the aspects measured (interest, unexpectedness and relevance) to be above three in the Likert scale. This is to ensure that this matches our working definition of serendipity ("making a pleasant and relevant discovery that was unexpected"), where pleasant is deduced from the fact that relevance and interest are both high.

- The estimation of the coping potential is based on the diversity of genres and formats in the user profile. We do not combine these diversity values, but we keep them separated to evaluate their importance. We divide users into two categories based on the diversity values: low and high coping potential.

The analysis has been performed in the following way:

1. **Comparison of the distributions of the similarity values**. We used a Wilcoxon Signed Rank test to identify differences in the similarity values based on the answers of the users: we divide the distributions of the answers in positive and negative (*i.e.*, higher than three (positive) and lower than three (negative) in the Likert scale) and we compare the distributions of the similarity values.

2. **Serendipity**. To verify whether the similarity values, together with the diversity of genre and format, are good predictors of serendipity, we perform a logistic regression analysis.

3. **Precision**. We calculate the precision of the recommendations by combining the assessments for interest and relevance (*i.e.*, considering the recommendation as liked when both interest and relevance assessments are higher than three in the Likert scale). To calculate the precision of the recommendations, we use the interest and the relevant answers (we did not ask directly to users if they like the TV programme recommended, as they are not actually watching it). First, we normalise the similarity values, then we apply the following formula:

$$\text{precision} = \frac{|\{\text{interesting progs}\} \cap \{\text{recommended progs}\}|}{|\{\text{recommended progrs}\}|} \quad (2)$$

where $|\{\text{interesting progs}\} \cap \{\text{recommended progrs}\}|$ is the number of recommendations evaluated positively (*i.e.*, ≥ 3) with similarity value ≥ 0.5 (*i.e.*, that the algorithm assessed as recommendable), and $|\{\text{recommended progrs}\}|$ is the number of programmes assessed as recommendable by the algorithm. The same reasoning has been applied for relevance and for relevance and interesting together.

4. **Catalog coverage**. We calculate the catalog coverage of the recommendations to verify the ability of the approaches to reach as many items in the catalog as possible.

Baseline

Comparison of the distributions of the similarity values.
We use the Wilcoxon Signed Rank test to compare the distributions of the similarity values when the answers are positive and negative. As we can see from Table 4, the rank of the distribution of the similarity values is low when interest is low. In the case of relevance, we see the same behaviour. While in the case of unexpectedness, we have a non-significant difference.

Null hypothesis	p-value
Positive interest $<$ Negative interest	**2.36e-06**
Positive relevance $<$ Negative relevance	**0.001121**
Positive unexpectedness $>$ Negative unexpectedness	0.9218

Table 4. Wilcoxon Signed Rank test result for values of similarity calculated with BBC Metadata. Positive indicates values higher than three in the Likert scale. Negative indicates values lower than three in the Likert scale.

Serendipity
The logistic regression results in a non-significant model for all the variables: similarity values, genre and format diversity. We report the model in Table 5,

Precision
The recommendations generated with BBC metadata reach a precision in terms of interest of 63%, in terms of relevance of 64%, and overall of 67%.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-3.5396	0.3522	-10.051	<2e-16
simValue	0.6914	0.8083	0.855	0.392
genreDiversity2	0.5493	0.3615	1.520	0.129
formatDiversity2	-0.2585	0.4588	-0.563	0.573

Table 5. Linear Regression Model with similarity values calculated with Metadata.

Approach	Precision interest	Precision relevance	Precision interest+relevance
BBC Metadata	0.63	0.64	0.67
LOD paths components	**0.68**	**0.69**	**0.71**
BBC Metadata & LOD paths components	0.67	0.65	0.69

Table 6. Precision of the three recommendation approaches.

Catalog coverage

The recommendations generated with BBC metadata reach a catalog coverage of 35,41%. In particular, the algorithm recommended 517 unique TV programmes, on 990 possible recommendations (6 recommendations per 165 users).

SIRUP

Comparison of the distributions of the similarity values.

We use the Wilcoxon Signed Rank test to compare the distributions of the similarity values when the answers are positive and negative. In Table 8, we can see that for interest the rank of the distribution of the similarity values is significantly higher when interest if high. We see a similar behaviour for relevance, while for unexpectedness we observe the opposite behaviour: the rank of the distribution of the similarity values is significantly lower when unexpectedness is high.

Serendipity

The logistic regression results in a model with significant variables the similarity values and the genre diversity, but not for the format diversity. In Table 9 we report the coefficients of the model. The results are interpreted as follows:

- for every one unit change in the simValue the log odds of a serendipitous recommendation (versus non-serendipitous) increases by 2.44.
- the indicator variable for genreDiversity have a slightly different interpretation. In particular, having a high coping potential (genrediversity2 indicate genre diversity above 5), versus having a low coping potential changes the log odds of finding a recommendation serendipitous by 0.7878.

Approach	Catalog Coverage
BBC Metadata	35,41%
LOD paths components	**47,40%**
BBC Metadata & LOD paths components	34,59%

Table 7. Catalog coverage of the three recommendation approaches.

Null hypothesis	p-value
Positive interest < Negative interest	**1.935e-07**
Positive relevance < Negative relevance	**1.404e-08**
Positive unexpectedness > Negative unexpectedness	**0.01096**

Table 8. Wilcoxon Signed Rank tests result for values of similarity calculated using LOD paths components. Positive indicates values higher than three in the Likert scale. Negative indicates values lower than three in the Likert scale.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-4.0018	0.4325	-9.252	<2e-16
simValue	2.4372	1.1480	2.123	0.0338
genreDiversity2	0.7878	0.3207	2.457	0.0140
formatDiversity2	0.1742	0.3478	0.501	0.6164

Table 9. Linear Regression Model with similarity values calculated with LOD paths components.

We can test for an overall effect of genre diversity by performing a Wald test. The results are: chi-squared test statistic of 9.8, with two degrees of freedom, is associated with a p-value of 0.0073 indicating that the overall effect of genre diversity is statistically significant. To ease the interpretation of the results, we exponentiate the coefficients of the model, results are shown in Table 10. We can now say that for one unit increase of the similarity value, the odds of having a serendipitous recommendation (versus a not serendipitous recommendation) increase by a factor of 11.44.

Precision

The recommendations generated with LOD path components reach a precision in terms of interest of 68%, in terms of relevance of 69%, and overall of 71%.

Catalog coverage

The recommendations generated with LOD path components reach a catalog coverage of 47,40%. In particular, the algorithm recommended 692 unique TV programmes, on 990 possible recommendations (6 recommendations per 165 users).

Combined approach

Comparison of the distributions of the similarity values.

We use the Wilcoxon Signed Rank test to compare the distributions of the similarity values when the answers are positive and negative. Table 11 shows that, in the case of interest, the rank of the distribution of the similarity values is lower when interest is higher. In the case of relevance, we see the opposite behaviour: the rank of the distribution of the similarity values is lower when relevance is low. Finally, for unexpectedness,

	Odds-Ratios	2.5 %	97.5 %
(Intercept)	0.01828252	0.007330145	0.03997674
simValue	11.44137704	1.356686051	121.16793171
genreDiversity2	2.19863554	1.173329612	4.15084088
formatDiversity2	1.19028314	0.588848705	2.31994835

Table 10. Odds ratios with a 95% confidence interval.

Null hypothesis	p-value
Positive interest > Negative interest	**0.0005365**
Positive relevance < Negative relevance	**0.020348**
Positive unexpectedness > Negative unexpectedness	**0.001782**

Table 11. Wilcoxon Signed Rank tests result for values of similarity calculated using LOD paths components and BBC metadata. Positive indicates values higher than three in the Likert scale. Negative indicates values lower than three in the Likert scale.

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	3.1971	0.3062	-10.442	<2e-16
simValue	-0.2566	0.9201	-0.279	0.780
genreDiversity2	0.5999	0.3549	1.690	0.091
formatDiversity2	0.2683	0.3406	0.788	0.431

Table 12. Linear Regression Model with similarity values calculated with LOD paths components and Metadata.

the rank of the distribution of the similarity values is lower when unexpectedness is higher.

Serendipity

The logistic regression results in a non-significant model for all the variables: similarity values, genre and format diversity. We report the model in Table 12.

Precision

The recommendations generated with BBC metadata and LOD path components reach a precision in terms of interest of 67%, in terms of relevance of 65%, and overall of 69%.

Catalog coverage

The recommendations generated with BBC metadata and LOD path components reach a catalog coverage of 34,59%. In particular, the algorithm recommended 505 unique TV programmes, on 990 possible recommendations (6 recommendations per 165 users).

DISCUSSION

We performed an experiment using three different approaches BBC metadata (Baseline), LOD paths components (SIRUP), BBC metadata and LOD paths components together (Combined approach) to generate recommendations and we analyse our results with three different analysis:

1. Comparison the distributions of the similarity values with a Wilcoxon Signed Rank test;
2. Serendipity with a logistic regression analysis;
3. Precision;
4. Catalog coverage.

The discussion of our findings will follow this division.

The **Comparison the distributions of the similarity values** give some preliminary insight into the performance of the algorithm with respect to the factors of serendipity. For the baseline, we see an expected behavior for relevance and interest (*e.g.*, when the similarity value is low, relevance and interest is low), see Table 4. The similarity values calculated with BBC metadata are not statistically different with respect to unexpectedness. For SIRUP, we also see the expected behavior for relevance and interest (Table 8). In this case, we also have a significant result for unexpected and, in particular, we see that higher the similarity, the lower the unexpectedness. This confirms the assumption that recommendations similar to what the user usually watch do not generate the surprise effect. Combining the two approaches, we can see that results show the same behavior as with SIRUP. It seems that merging the approaches result in weaker results for interest and relevance, but a stronger result for unexpectedness. However, in the latter case, the difference is not significant. So we can conclude that SIRUP performs better.

The logistic regression analysis to model **serendipity** seems to confirm the previous findings. In particular, we have not-significant models with the baseline and the combined approach, while we have a significant model with SIRUP. We included in the model also the estimation of the coping potential with genre and format diversity in the user profile. Our findings are interesting: not only the similarity values calculated with LOD paths components is a significant variable in the model (see Table 9), but also the genre diversity, treated as a categorical variable is significant in the model: if the user has a high genre diversity in her profile, she is more open to serendipitous recommendations. This is not the case when looking at the format diversity, which is a non-significant variable in the model. We believe that this is due to the fact that the format is not a good discriminant for topics, while the genre is. Another interesting fact to notice is that the similarity values are a positive variable in the model, meaning that when it grows by a unit, the odds ratios of a serendipitous recommendation increase by a factor of 11.44 (see Table 10).

These results support the following hypotheses:

RQ2: *Can we perform the novelty check of TV programmes with respect to the user profile using LOD paths components?*

RQ3: *Can we estimate the coping potential of a user with the diversity of the TV programmes in the user profile?*

It is known that aiming for serendipity can be harmful to **precision** [24]. We found that on average, SIRUP reaches 71% precision, which is higher both than the baseline and the combined approach. Also the catalog coverage is higher with SIRUP (47,40%), than with the other approaches.

Overall, our findings support also our first research question:

RQ1: *Do serendipitous recommendations trigger curiosity in users?*

Our findings allow us to conclude that curiosity and serendipity are two connected concepts, and that using curiosity theory as guidance to model serendipitous recommendations is a promising approach.

CONCLUSION

In this paper, we presented a model for serendipity in content-based recommender system (SIRUP) inspired by curiosity theories. We show that the two elements that compose SIRUP

(novelty check and coping potential check) are essential aspects of the serendipitous assessment. Besides the verification of the model, we also show that our approach with LOD paths allows us to perform properly the novelty check in order to identify the serendipitous recommendations. We were also able to show that the coping potential check is an important aspect in the process of identifying the right level of serendipity for different users. We show that a good estimator for the coping potential of the users is the genre diversity of the TV programmes in her profile.

As future work, we want to deepen the study of the coping potential, by including a short questionnaire about the user attitude towards new knowledge. This is mainly due to the fact that when building user profiles, the observation of the user behaviour is always partial. With a questionnaire, we should be able to overcome this limitation. We also want to improve the assessment of serendipity, through implicit feedback, as done by de Gemmis et al. [8]. We believe that this approach can be successful also in other contexts, like books and culture heritage. We aim at taking full advantage of the semantic enrichment, by extracting LOD paths also for concepts in the other textual metadata (*i.e.*, genre, credits, and so on).

ACKNOWLEDGMENT

This research was supported by the EU FP7 STREP "ViSTA-TV" project and by the Network Institute "SIRUP" project. We would like to thank our colleagues Allison Eden, Tilo Hartmann and Britt Hoeksema for their support and help.

REFERENCES

1. Akiyama, T., Obara, K., and Tanizaki, M. Proposal and evaluation of serendipitous recommendation method using general unexpectedness. In *Workshop on the Practical Use of Recommender Systems, Algorithms and Technologies (PRSAT 2010)*, CEUR-WS.org (2010), 3.
2. Berlyne, D. E. A theory of human curiosity. *British Journal of Psychology. General Section 45*, 3 (1954), 180–191.
3. Berlyne, D. E. *Conflict, arousal, and curiosity.* McGraw-Hill Book Company, New York, NY, US, 1960.
4. CBS - Statistic Netherlands. Statistics Netherlands: 9 in 10 people access the internet every day [Press release]. Retrieved from `http://goo.gl/hjjqrp`, 2015.
5. Cremonesi, P., Garzotto, F., and Turrin, R. Investigating the persuasion potential of recommender systems from a quality perspective: An empirical study. *ACM Trans. Interact. Intell. Syst. 2*, 2 (June 2012), 11:1–11:41.
6. Cremonesi, P., Garzottto, F., and Turrin, R. User effort vs. accuracy in rating-based elicitation. In *Proceedings of the Sixth ACM Conference on Recommender Systems*, RecSys '12, ACM (New York, NY, USA, 2012), 27–34.
7. Csikszentmihalyi, M. Flow: The psychology of optimal experience, 1990.
8. de Gemmis, M., Lops, P., Semeraro, G., and Musto, C. An Investigation on the Serendipity Problem in Recommender Systems. *Inf. Process. Manage. 51*, 5 (Sept. 2015), 695–717.
9. De Vocht, L., Beecks, C., Verborgh, R., Mannens, E., Seidl, T., and Van de Walle, R. Effect of heuristics on serendipity in path-based storytelling with linked data. In *Human Interface and the Management of Information:Information, Design and Interaction: 18th International Conference, HCI International 2016 Toronto, Canada, July 17-22, 2016, Proceedings, Part I*, S. Yamamoto, Ed., Springer International Publishing (Cham, 2016), 238–251.
10. Dijkshoorn, C., Aroyo, L., Schreiber, G., Wielemaker, J., and Jongma, L. Using Linked Data to Diversify Search Results a Case Study in Cultural Heritage. In *Proceedings of the 19th International Conference on Knowledge Engineering and Knowledge Management*, Springer International Publishing (Cham, 2014), 109–120.
11. Dimitrova, V., Lau, L., Thakker, D., Yang-Turner, F., and Despotakis, D. Exploring exploratory search: A user study with linked semantic data. In *Proceedings of the 2Nd International Workshop on Intelligent Exploration of Semantic Data*, IESD '13, ACM (New York, NY, USA, 2013), 2:1–2:8.
12. Dojchinovski, M., and Vitvar, T. Personalised access to linked data. In *Knowledge Engineering and Knowledge Management: 19th International Conference, EKAW 2014, Linköping, Sweden, November 24-28, 2014. Proceedings*, Springer International Publishing (Cham, 2014), 121–136.
13. Ericsson Consumerlab. TV AND MEDIA 2015. The empowered TV and media consumer's influence. Tech. rep., Ericsson, 2015. Available online at `http://goo.gl/0iEWDv`.
14. Gaffney, T. If Content is King, Video is Heir to the Throne. `http://goo.gl/Q79j2L`, 2013.
15. Hartig, O. Sparql for a web of linked data: Semantics and computability. In *The Semantic Web: Research and Applications: 9th Extended Semantic Web Conference, ESWC 2012, Heraklion, Crete, Greece, May 27-31, 2012. Proceedings*, Springer Berlin Heidelberg (Berlin, Heidelberg, 2012), 8–23.
16. Hollink, L., Malaisé, V., and Schreiber, G. Enriching a thesaurus to improve retrieval of audiovisual documents. In *Semantic Multimedia: Third International Conference on Semantic and Digital Media Technologies, SAMT 2008, Koblenz, Germany, December 3-5, 2008. Proceedings*, Springer Berlin Heidelberg (Berlin, Heidelberg, 2008), 47–60.
17. Iaquinta, L., de Gemmis, M., Lops, P., Semeraro, G., Filannino, M., and Molino, P. Introducing serendipity in a content-based recommender system. In *Proceedings of the 2008 8th International Conference on Hybrid Intelligent Systems*, HIS '08, IEEE Computer Society (Washington, DC, USA, 2008), 168–173.

18. Kashdan, T. B. Curiosity. In *Character strengths and virtues: A handbook and classification*, C. Peterson and M. E. P. Seligman, Eds. Oxford University Press and American Psychological Association, New York and Washington DC, 2004, 125–141.

19. Loewenstein, G. The psychology of curiosity: A review and reinterpretation. *Psychological Bulletin 116*, 1 (1994), 75–98.

20. Maccatrozzo, V., Aroyo, L., and Van Hage, W. R. Crowdsourced evaluation of semantic patterns for recommendation. In *Late-Breaking Results, Project Papers and Workshop Proceedings of the 21st Conference on User Modeling, Adaptation, and Personalization., Rome, Italy, June 10-14, 2013*, vol. 997 of *CEUR Workshop Proceedings*, CEUR-WS.org (2013), 15–21.

21. Macedo, P., Cardoso, J., and Pinto, A. M. Enriching Electronic Programming Guides with Web Data. In *Proceeding of the 2nd International Workshop on Linked Media (LiME2014)* (Crete, Greece, 2014).

22. Malaisé, V., Isaac, A., Gazendam, L., and Brugman, H. Anchoring dutch cultural heritage thesauri to wordnet: Two case studies. In *Proceedings of the Workshop on Language Technology for Cultural Heritage Data (LaTeCH 2007)*, ACL (Prague, Czech Republic, 2007), 57–64.

23. Marie, N., Gandon, F., Ribière, M., and Rodio, F. Discovery hub: On-the-fly linked data exploratory search. In *Proceedings of the 9th International Conference on Semantic Systems*, I-SEMANTICS '13, ACM (New York, NY, USA, 2013), 17–24.

24. McNee, S. M., Riedl, J., and Konstan, J. A. Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *CHI '06 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '06, ACM (New York, NY, USA, 2006), 1097–1101.

25. Musto, C., Narducci, F., Lops, P., Semeraro, G., de Gemmis, M., Barbieri, M., Korst, J. H. M., Pronk, V., and Clout, R. Enhanced semantic tv-show representation for personalized electronic program guides. In *User Modeling, Adaptation, and Personalization: 20th International Conference, UMAP 2012, Montreal, Canada, July 16-20, 2012. Proceedings*, Springer Berlin Heidelberg (Berlin, Heidelberg, 2012), 188–199.

26. Ossenbruggen, J., Amin, A., Hardman, L., Hildebrand, M., Assem, M., Omelayenko, B., Schreiber, G., Tordai, A., de Boer, V., Wielinga, B., Wielemaker, J., Niet, M., Taekema, J., Orsouw, M. F., Teesing, A., Trant, J., and Bearman, D. Searching and annotating virtual heritage collections with semantic-web techniques. In *Proceedings of Museums and the Web 2007*, Archives & Museum Informatics (March 2007), 1 – 11. Available at `http://goo.gl/jJGVU5`.

27. Pariser, E. *The Filter Bubble: What the Internet Is Hiding from You*. Penguin Group, The, New York, NY, 2011.

28. Saleem, M., Kamdar, M. R., Iqbal, A., Sampath, S., Deus, H. F., and Ngonga Ngomo, A.-C. Fostering serendipity through big linked data. In *Semantic Web Challenge at ISWC2013* (2013).

29. Schein, A. I., Popescul, A., Ungar, L. H., and Pennock, D. M. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, ACM (New York, NY, USA, 2002), 253–260.

30. Shadbolt, N., Berners-Lee, T., and Hall, W. The semantic web revisited. *IEEE Intelligent Systems 21*, 3 (Jan 2006), 96–101.

31. Silvia, P. J. Emotional responses to art: From collation and arousal to cognition and emotion. *Review of General Psychology 9* (2005), 342–357.

32. Silvia, P. J. Interest - the curious emotion. *Current Directions in Psychological Science 17*, 1 (2008), 57–60.

33. Waitelonis, J., and Sack, H. Towards exploratory video search using linked data. *Multimedia Tools and Applications 59*, 2 (2012), 645–672.

34. Walpole, H. To Mann, Monday 18 January 1754. In *Horace Walpole's Correspondence*, W.S. Lewis, Ed., vol. 20. Yale University Press, New Haven, Connecticut, USA, 1960, 407–411.

35. Zhang, Y. C., Séaghdha, D., Quercia, D., and Jambor, T. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining*, WSDM '12, ACM (New York, NY, USA, 2012), 13–22.