

# Effect of Item Representation and Item Comparison Models on Metrics for Surprise in Recommender Systems

Andre Paulino de Lima<sup>a</sup> and Sarajane Marques Peres<sup>b</sup>

*School of Arts, Sciences and Humanities, University of São Paulo, Brazil*

**Keywords:** Recommender Systems, Surprise Metric, Unexpectedness, Serendipity, Item Representation, Item Comparison, Off-line Evaluation.


**Abstract:** Surprise is a property of recommender systems that has been receiving increasing attention owing to its links to serendipity. Most of the metrics for surprise poorly agree with definitions employed in research areas that conceptualise surprise as a human factor, and because of this, their use in the task of evaluating recommendations may not produce the desired effect. We argue that metrics with the characteristics that are presumed by models of surprise from the Cognitive Science may be more successful in that task. Moreover, we show that a metric for surprise is sensitive to the choices of how items are represented and compared by the recommender. In this paper, we review metrics for surprise in recommender systems, and analyse to which extent they align to two competing cognitive models of surprise. For that metric with the highest agreement, we conducted an off-line experiment to estimate the effect exerted on surprise by choices of item representation and comparison. We explore 56 recommenders that vary in recommendation algorithms, and item representation and comparison. The results show a large interaction between item representation and item comparison, which suggests that new distance functions can be explored to promote serendipity in recommendations.


## 1 INTRODUCTION

The purpose of a Recommender System (RS) is to enable its user to select an item within a universe of items that is predominantly unknown by them. A recommender can suggest to its users any of the items in its repository, and such items can represent anything of interest to users, such as books, movies, music, hotels, restaurants, or scientific articles. In its most rudimentary form, a recommender produces non-personalised recommendations by offering items of popular interest to all users. In this case, an RS uses descriptive statistics of the ratings given to the items by the users to produce recommendations that consist of items with higher expected rating. However, numerous approaches have been proposed over time to allow an RS to produce personalised recommendations by making use of different sources of information. To select the personalisation approach that is the most appropriate for a given domain (e.g. movie or music recommendation), it is necessary to define an evaluation method that can be used to compare

the performance of different approaches and to determine which one makes the recommender more or less competent in that domain. It must be noted that the evaluation method can assist the owner of the system not only in selecting a personalisation approach but also in tuning a recommender instance, assuming that owner of the system has a method for seeking optimal parameters according to the objectives imposed by the application. It must be said that, for a long time, the most common property associated to the competence of a recommender was its predictive accuracy.

As the research area evolved, a consensus on the inadequacy of adopting accuracy as the only relevant property for an application emerged among researchers (Herlocker et al., 2004; McNeer et al., 2006). Nowadays, beyond-accuracy properties of recommender systems is known to play a critical role in user satisfaction and, among these properties, surprise has recently been the subject of several studies owing to its links to serendipity (Adamopoulos and Tuzhilin, 2011; Kaminskis and Bridge, 2014; Silveira et al., 2017) and the problem of over-specialisation in content-based recommenders (de Gemmis et al., 2015), as well as its importance in some application domains (Mourão et al., 2017). The notion of surprise

<sup>a</sup>  <https://orcid.org/0000-0002-3148-6686>

<sup>b</sup>  <https://orcid.org/0000-0003-3551-6480>

generally reflects the capacity to make recommendations that are dissimilar from the items known to a given user (Kaminskas and Bridge, 2014; Adamopoulos and Tuzhilin, 2011; Zhang et al., 2012).

However, the metrics in the literature show a low level of conceptual agreement with models of surprise being investigated in Cognitive Science, in which surprise is framed as a human factor. These metrics explore intuitions about how to assess the surprise experienced by a user when they are exposed to a recommendation list. They differ in terms of how much they can meet requirements imposed by the subjectivity inherent in the perceptions of surprise, by the dynamism of the perception of a user who is constantly exposed to new experiences, and by the difficulty of evaluating the discrepancy (or distance) between what is expected and what is observed. Analysing the relationship of measures with such capacities is the first step for those who intend to understand them. The next step should be exploring the choice of models for representing and comparing items, for instance, with a view to evaluating a recommendation list in a recommender system (or to another application domain of particular interest). Such choices are expected to affect the performance of metrics. Since the knowledge of which combinations of item representations, item comparison and, in our case, recommender algorithms have a greater or lesser impact on such metrics generally comes from costly empirical experiments. The documentation of results from experiments of this nature is beneficial for the acceleration of future studies.

In this paper we present the effects of choices related to item representation and item comparison models on the results obtained by applying metrics for surprise, considering combinations of: four item representation models (count-based and prediction-based distributional semantics models, sparse and factorised user-item models); six distance functions to implement item comparison models (Euclidean and cosine distances as geometric intuitions, Jaccard distance as a combinatorial intuition, Kullback-Leibler and Jensen-Shannon divergences as information intuitions, and Aitchison distance as a statistical intuition); and four algorithms to create recommendation lists (factorisation and three variations of k-nearest neighbours). The context of movies recommendation was chosen to support the systematic experiment that allowed the analysis of the effects. Thus, as contributions of this work we stand out:

1. proposition of an evaluation framework to support analysis on the level of agreement of each metric with two competing cognitive models of surprise: the cognitive-evolutionary model (Meyer et al.,

1997) and the metacognitive explanation based model (Foster and Keane, 2015). This framework is introduced in section 3;

2. organisation and public availability of an extended dataset which complements the MovieLens-1M Dataset (Harper and Konstan, 2015) by explicitly associating movies with their respective textual short description obtained from online MovieLens system. This extended dataset offers a richer experimentation environment to the academic community since it facilitates the conduction of experiments with item content. Moreover, the public availability improves the reproducibility conditions related to the discussions developed in this paper.
3. investigation on how a metric for surprise is affected by choices of item representation (how items are mapped to numerical representations) and item comparison (what is the notion of similarity that is captured), through a systematic experimentation that tests 56 recommenders executed with the extended MovieLens dataset.

This paper is organised as follows: Section 2 looks into the metrics of surprise proposed for recommender systems in the last decade, and the state-of-the-art in off-line evaluation method for surprise. Section 3 presents an analysis of the extent to which each metric is in agreement with cognitive models of surprise following a framework introduced in this paper and justifies the adoption in our experiment of the metric that achieves the highest agreement. Section 4 describes the results obtained from an experiment conducted to estimate the effect of item representation and item comparison on the surprise of a recommender system. Section 5 presents our conclusions.

## 2 BACKGROUND

This section begins by addressing the relationship between surprise and other desired properties of recommender systems, and a terminology issue that this relation raises. Then, a review of metrics for surprise is presented. The section ends with a brief description of the *one plus random* evaluation method, with which the metrics can be used and tested.

### 2.1 The Property of Surprise

The challenge of discovering new items that might be useful to a user has been the focus of a many works in the literature on recommender systems. In general, the approach involves finding new items that bear

some similarity to items which have been given good ratings by a set of selected users. An even greater challenge is to find new items that do not resemble items known to a user, yet would still be useful to them. This would be a serendipitous recommendation. In Herlocker et al. (2002), the authors introduced a definition of serendipity now widely cited: “A serendipitous recommendation helps the user find a surprisingly interesting item he might not have otherwise discovered.” In a sense, this definition supports a perspective whereby serendipity, as a system property, results from the interaction of two other and more fundamental properties: surprise and relevance. In this view, being surprising and relevant (or useful) to a user are the basic requirements of a serendipitous recommendation.

It has been recently pointed out in Kaminskas and Bridge (2016) that there is a conceptual overlap between the properties of novelty or unexpectedness and the notion of surprise<sup>3</sup>. Thus, the same notion may be associated with different names in the literature. In this study, we subscribe to the categorisation suggested by Kaminskas and Bridge (2016), in which (a) novelty is related to the notion of an item being popular, and thus is not directly related to serendipity, (b) unexpectedness usually conveys the same notion as surprise, and (c) surprise can be regarded as a component of serendipity. In view of this, our focus is on the metrics for estimating surprise, and the metrics for serendipity or unexpectedness whose definitions involve the notion of surprise. We claim that the metrics in the literature, although clearly distinct from each other, all address some factors commonly related to scale construction. We select two of these factors to serve as contrasts in our analysis:

- *Intrinsic vs extrinsic evaluation*: some metrics only use data that are internal to the system under evaluation (Akiyama et al., 2010; Zhang et al., 2012; Kaminskas and Bridge, 2014); other metrics use data made available by an external system<sup>4</sup> in addition to data that is internal to the system under evaluation (Murakami et al., 2008; Ge et al., 2010; Adamopoulos and Tuzhilin, 2011).
- *Subjective vs objective view*: some metrics assume that surprise is subjective in nature, since it depends on the user past experience, which usually is represented by the set of items known to a user (Murakami et al., 2008; Adamopoulos and Tuzhilin, 2011; Zhang et al., 2012; Kaminskas

and Bridge, 2014); other metrics view surprise as a property of the item (Ge et al., 2010; Akiyama et al., 2010) and, thus, is independent of the users.

## 2.2 Metrics for Surprise

In this section, we review six surprise-related metrics. All of them involve some notion of distance, but they do not agree on the subjectivity of surprise, neither on the necessity of the information used in the assessment to be internal to the system being evaluated. Figure 1 illustrates how they are positioned with regard to these two factors. In the figure, the metrics are presented by year of publication, and the ellipses show trends or changes in the factors. This review is not meant to be exhaustive but rather aims to capture the main approaches that have evolved over the years.

### 2.2.1 A Metric for Unexpectedness

In Murakami et al. (2008), it was proposed a metric to evaluate serendipity. It relies on the ideas that (a) recommendations made by a PPM are prone to be obvious, and (b) a serendipitous recommendation must be non-obvious. The metric is calculated from a recommendation list  $L$  produced for a user  $u$  by the system being evaluated (Equation 1).

The predicate<sup>5</sup>  $rscore$  accounts for the predicted relevance of an item  $L_i$  to the user  $u$ , while  $isrel$  accounts for surprise, and reflects the degree to which an item  $L_i$  is similar to items highly rated by the user (a subjective view).

The metric performs an extrinsic evaluation because  $rscore$  (Equation 2) combines the relevance predicted by the system under evaluation ( $Pr$ ) with the relevance predicted by an external system ( $Pr^*$ ).

$$unexp(L, u) = \frac{1}{|L|} \sum_{i=1}^{|L|} rscore(L_i, u) \times isrel(L_i, u) \quad (1)$$

$$rscore(L_i, u) = \max(Pr(L_i, u) - Pr^*(L_i, u), 0). \quad (2)$$

### 2.2.2 A Metric for Serendipity

In Ge et al. (2010), another metric was devised to assess serendipity. As shown in Equation 3,  $srdp$  is applied to a list  $L^\delta$ , and estimates the average *usefulness* of its items,  $L_i^\delta$ . In Equation 4,  $L^\delta$  is obtained from the difference between the list  $L$ , generated by the system under evaluation for the user  $u$ , and the list  $L^*$ , drawn up for user  $u$  by an external system. This means that

<sup>5</sup>The term *predicate* denotes a procedure that performs some computation. It is similar to *function* except that its domain and codomain are implicit, and thus no assurance can be given about its success in completing a computation.

<sup>3</sup>A similar overlap has been pointed out by Barto et al. (2013) in the literature on cognitive science.

<sup>4</sup>Such a system is often referred to as baseline system or PPM - Primitive Prediction Model.

		2007	2010		2011	2012	2014
		Murakami et al.	Ge et al.	Akiyama et al.	Adamopoulos and Tuzhilin	Zhang et al.	Kaminskas and Bridge
Evaluation	intrinsic						
	extrinsic						
View	subjective						
	objective						

Figure 1: Evolution of surprise-related metrics. A metric is classified as intrinsic if the data it uses are exclusively internal to the system being evaluated, and it is classified as subjective if it explicitly identifies the user experience in its definition.

$L^\delta$  comprises non-obvious, unexpected items, and accounts for surprise. Thus,  $srdp$  performs an extrinsic evaluation. In addition, we argue that it operates in an objective way, since the user experience is not considered when surprise is assessed.

$$srdp(L^\delta, u) = \frac{1}{|L^\delta|} \sum_{i=1}^{|L^\delta|} usefulness(L_i^\delta, u) \quad (3)$$

$$L^\delta = L \setminus L^* \quad (4)$$

### 2.2.3 A Metric for General Unexpectedness

In Akiyama et al. (2010), it was set out a metric called “general unexpectedness” that explores a combinatorial intuition: an item that shows a rare combination of attributes must be taken as unexpected. It assumes that each item has some content combined with it, in the form of a set of attributes. This usually is the case with content-based recommenders (de Gemmis et al., 2015). As shown in Equation 5, the  $unexp$  metric is estimated for  $L$ , the recommendation list produced to user  $u$  by the system being evaluated, and this aggregates the  $uscore$  obtained for each item  $L_i$ . The  $uscore$ , defined in Equation 6, is the reciprocal of the joint probability estimated for each possible pair of attributes of  $L_i$ . In this equation,  $A(L_i)$  represents the set of attributes that describe  $L_i$ ,  $N_a$  denotes the number of items in the repository that have attribute  $a$  and  $N_{a,b}$  is the number of items that have both attributes  $a$  and  $b$ . Thus an objective view is adopted since surprise can be seen as a property of the content of an item. Unlike the previously described metrics, this one does not employ an external system, and, thus, performs an intrinsic evaluation.

$$unexp(L) = \frac{1}{|L|} \sum_{i=1}^{|L|} uscore(L_i) \quad (5)$$

$$uscore(L_i) = \left[ \frac{1}{|A(L_i)|} \sum_{a,b \in A(L_i)} \frac{N_{a,b}}{N_a + N_b - N_{a,b}} \right]^{-1} \quad (6)$$

### 2.2.4 A New Metric for Unexpectedness

In Adamopoulos and Tuzhilin (2011), the proposed metric explores an intuition about user expectation: an item is expected by a user if it can be anticipated. The  $unexp$  metric (Equation 7) is calculated from  $L$ , the recommendation list produced for user  $u$ , and  $L^s$ , a list of obvious, expected items (Equation 8). The recommendation list  $L^*$  is produced for user  $u$  by an external system,  $E_u$  represents the set of items that have been rated by user  $u$ , and the predicate  $neighbours$  represents the set of items in the system repository  $I$  that are similar to the items in  $E_u$  up to some degree specified by threshold parameters in  $\theta$ . Thus, this metric performs an extrinsic evaluation, and subscribes to the subjective view.

$$unexp(L, L^s) = \frac{1}{|L|} |L \setminus L^s| \quad (7)$$

$$L^s(u) = L^* \cup E_u \cup neighbours(I, E_u, \theta) \quad (8)$$

### 2.2.5 The Unserendipity Metric

In Zhang et al. (2012), it was considered that a serendipitous recommendation must be dissimilar to items known to the user, in a semantic sense. It resembles the metric proposed by Akiyama et al. (2010), since it assumes that each item is combined with some content, but in this case, this content is organised as vectors in  $\mathbb{R}^m$ . The metric is computed from the recommendation list  $L$  drawn up for user  $u$  (Equation 9), and results in a score that is the average cosine similarity obtained from the items in  $L$  and the set of items known to the user,  $E_u$ . It does not use an external system (intrinsic evaluation), and it adheres to a subjective view of surprise. It must be noted that the metric



is scale-inverted: the lower the score, the more surprising  $L$  is.

$$unsrdp(L, u) = \frac{1}{|L||E_u|} \sum_{i \in L} \sum_{j \in E_u} \text{cossim}(i, j) \quad (9)$$

### 2.2.6 A Metric for Surprise

In a similar way to Zhang et al. (2012), in Kaminskias and Bridge (2014), it was argued that a surprising recommendation must be dissimilar to items known to the user, but does not require that this dissimilarity should be semantic in nature. They also explore the interplay between the notions of distance and similarity<sup>6</sup>. The metric is calculated from the recommendation list  $L$  produced for user  $u$  (Equation 10) and assesses the average surprise obtained for each item in  $L$ . The surprise of an item  $i$  in  $L$  is estimated as the minimum distance between  $i$  and the set  $E_u$  containing all items known to the user (Equation 11), or as the maximum degree of similarity between  $i$  and  $E_u$  (Equation 12). The function  $dist$  is defined as the Jaccard distance between the set of attributes extracted from contents linked to items  $i$  and  $j$ , and  $sim$  computes the normalised pointwise mutual information score (NPMI) (Bouma, 2009) for the same items. This metric does not use an external system (intrinsic evaluation), and supports a subjective view of surprise, since it takes account of the user's experience.

$$surprise(L, u) = \frac{1}{|L|} \sum_{i \in L} S_i(i, E_u) \quad (10)$$

$$S_i(i, E_u) = \min_{j \in E_u} dist(i, j) \quad (11)$$

$$S_i(i, E_u) = \max_{j \in E_u} sim(i, j) \quad (12)$$

## 2.3 Evaluation Method for Surprise

All the metrics described in Section 2.2 evaluate a single recommendation list. Thus, an evaluation method is required to obtain an estimate of how the system performs with regard to surprise. Most studies follow a statistical procedure: a sample of users is randomly drawn, recommendation lists are produced to those users, surprise evaluations are made, and the average surprise is taken as the desired estimate.

To estimate the recall property of a recommender instance in a top-N recommendation task, the off-line evaluation method *one plus random* can be adopted (Cremonesi et al., 2008; Bellogin et al., 2011). This method follows the intuition that, in a sufficiently

<sup>6</sup>Given a distance function, a similarity can be derived over the same domain (Deza and Deza, 2009).

large set  $L_1$  that consists of items unknown to user  $u$ , most of these items are irrelevant to  $u$ .

In Kaminskias and Bridge (2014) this method was adapted to estimate the degree of surprise of a recommender system. The original intuition of the method is retained and, in addition, to computing an estimate for recall, it also computes the average surprise obtained from the recommendation lists produced for a sample of users.

## 3 THE PROPERTY OF SURPRISE REVISITED

As seen in the previous section, metrics for surprise explore diverse intuitions to model the relationship between the surprise experienced by a user and the recommendation list that was presented to them. In fact, we argue that a metric for surprise would benefit from, and should account for, the ideas that have been explored in the field of Cognitive Science about how humans experience surprise.

Reisenzein et al. (2017) examined the extent to which the experimental evidence supports current models of surprise, in particular: the cognitive-evolutionary model (Meyer et al., 1997), and the metacognitive explanation-based model (Foster and Keane, 2015)<sup>7</sup>.

According to the cognitive-evolutionary model, the feeling of surprise emerges as a response to “unexpected (schema-discrepant) events” that convey a change in the environment. In contrast, the metacognitive explanation-based model approaches surprise as a response to a failure to track the cause of a change in the environment. However, despite of any divergences, the two theories converge about the subjective nature of surprise, and, to some extent, both models are aligned with the definition of surprise as a distance, as has been proposed in Itti and Baldi (2009).

As the first contribution of this paper, we propose a simple framework to assess the degree of coherence between a metric for surprise and the set of characteristics presumed by the cognitive models, and apply it to the metrics for surprise reviewed in the previous section. The result of this assessment is summarised in Table 1, and, by adopting this framework, we argue that the metric proposed by Kaminskias and Bridge (2014) is in higher agreement to the cognitive models of surprise because:

- (a) it adopts a subjective view, unlike Ge et al. (2010); Akiyama et al. (2010), and it does not depend on

<sup>7</sup>This study reviews several models, but these two models are particularly useful to this analysis.

Table 1: Metrics for surprise in recommender systems and their adherence to characteristics presumed by the two cognitive models. A metric is adherent to the notion of subjectivity if it considers that surprise not only depends on the user experience (column “is subjective”), but also that this experience must be exclusively internal to the user (column “is intrinsic”).

Metric	Subjectivity		Sensitivity to user experience
	Is subjective?	Is intrinsic?	Is dynamic?
(Ge et al., 2010)	no	—	—
(Akiyama et al., 2010)	no	—	—
(Murakami et al., 2008)	yes	no	—
(Adamopoulos and Tuzhilin, 2011)	yes	no	—
(Zhang et al., 2012)	yes	yes	lower
(Kaminskas and Bridge, 2014)	yes	yes	higher

information that resides externally to the system being evaluated, unlike Murakami et al. (2008); Adamopoulos and Tuzhilin (2011);

- (b) it accounts for changes in the surprise of an unobserved item, as the growth of the set of items known to the user ( $E_u$ ), and is more sensitive in this regard than the metric proposed by Zhang et al. (2012);
- (c) it is proportional to the degree of dissimilarity between a recommended item  $i$  and the items known to the user, which embeds notions of distance.

However, this metric pursues a naïve intuition about surprise and only gives a superficial account of the real user experience. For example, it can be seen as an exhaustive search in the user memory  $E_u$  for similar events. From this perspective, it assumes that every individual can recall every past event equally well, and thus fails to account for known cognitive biases, such as recall and retrievability biases. The former refers to the relative ease to recall recent and vivid events in relation to events that were observed in a remote past or were unemotional (Tversky and Kahneman, 1974; Bazerman and Moore, 2009). The latter refers to how the subjective context can modify the relative salience to our perception of the features of an object, and the role that this salience plays in our judgment of similarity (Tversky, 1977; Gershman, 2017).

Despite of these limitations, and in the absence of metrics of higher fidelity, we argue that this metric is still useful to estimate surprise in recommender systems, and we adopt it in our experiment.

## 4 EXPERIMENTAL RESULTS

The second and third contributions of this paper refer respectively to the extended dataset and to the systematisation of an experiment that aim to analyse

the effects of item representation and item comparison models on the measures for surprise in a recommender system. Both are presented in this section.

### 4.1 Experiment Setup

This section presents the dataset, method, models and algorithms used in the experiment.

#### 4.1.1 Dataset

The MovieLens-1M Dataset (Harper and Konstan, 2015) was used to build the repository of the recommender system instance. The dataset contains 3,883 items (title of movies and a set of genres in which each movie fits), 6,040 users (demographic data) and just over 1 million ratings (tuples containing a user, a movie, the rating given by the user to the movie and a temporal reference to when the rating was created). To investigate the effect of item representation in the metrics for surprise, the dataset was extended to combine a short textual description to each movie.

This extension enables the adoption of data representation models that rely on textual content to produce item vectors. Thus, we enhanced the MovieLens-1M dataset by combining a short textual description to each item. These short descriptions were collected from the online MovieLens system in September 2017. Items whose description was not available, too short, or not written in English were rejected, as well as items with no rating. These constraints aim at controlling the variability between our experiment specifications at the price of a small reduction in the number of items and ratings: 3,643 and 997,136, respectively. The extended dataset adopted the same format that is employed by the MovieLens datasets<sup>8</sup>.

<sup>8</sup>The code and dataset can be downloaded from <https://github.com/andreplima/surprise-in-recsys>.

#### 4.1.2 Method

A controlled environment was created<sup>6</sup>. It has three components: the experiment specification (config), a recommender system instance, and an evaluator. The config specifies: (a) a recommendation algorithm, (b) an item representation scheme (i.e. how item vectors are created), (c) an item comparison scheme (i.e. a distance function that can be applied to item vectors), and (d) a sample of users. The recommender system produces a single recommendation list to each user in the sample by means of the recommendation algorithm specified in the config. The evaluator assesses surprise in the lists by means of a surprise metric (Section 3) and the *one plus random* method (Section 2.3). The metric always uses the item representation and distance function specified in the config. The sample of users was drawn once and reused across all configs, and its size ( $n = 362$ ) was selected to obtain a mean surprise estimate with a 5% error margin and 95% confidence.

Given a target config, the evaluator assesses the average degree of surprise of the recommendation lists produced by the recommender system to the users in the sample. The surprise obtained from each recommendation list is collected, and the system-level estimate is taken as the observed average. As detailed next, each config specifies one of four item representation schemes, one of six distance functions, and one of four recommendation algorithms.

#### 4.1.3 Item Representation

We selected four models, including models employed in content-based and collaborative approaches: Models C and D are distributional semantics models (DSM) based on the document-term matrix, and Models U and V are based on the user-item matrix:

- *Model C (Count-based DSM)* is a vector space model of semantics (Baroni et al., 2014; Turney and Pantel, 2010). It uses the short textual description linked to the items to produce a document-term matrix, from which the item vectors are extracted. Tokens were stemmed by means of the Snowball algorithm (Porter, 2001; Loper and Bird, 2002) before computing tf-idf scores for each short description (Manning et al., 2008). Since this is a sparse model, item vectors have 13,797 dimensions (the number of terms in the corpus vocabulary).
- *Model D (Predictive DSM)* is another vector space model of semantics. The Paragraph Vector algorithm (Mikolov et al., 2013; Řehůřek and Sojka, 2010) is applied on the short descriptions to

extract item vectors. This model can be seen as a factorised model induced from the document-term matrix (Levy and Goldberg, 2014). Item vectors with 100 dimensions were extracted.

- *Model U (Sparse UI)* is a sparse user-item model (Ning et al., 2015). Each item  $i$  is represented as a vector  $\mathbf{r}_i$  of length equal to the number of users in the repository (6,040 users). A vector element  $r_{ui}$  is taken as either the rating the user  $u$  has attributed to item  $i$ , or zero if the item  $i$  was not rated by the user  $u$ .
- *Model V (Factorised UI)* is a factorised user-item model (Ning et al., 2015). Each item  $i$  is a column vector  $\mathbf{q}_i \in \mathbf{Q}$ , which is obtained by the PureSVD algorithm, that is detailed ahead. Item vectors with 100 dimensions were extracted.

#### 4.1.4 Item Comparison

Six distance functions were selected. They explore distinct intuitions about separation when comparing item vectors: Euclidean and cosine distances (geometric intuition), the Jaccard distance (combinatorial), the Kullback-Leibler and Jensen-Shannon divergences (informational), and the Aitchison distance (Egozcue et al., 2011) (statistical).

However, their application to item vectors from some representation models is hindered owing to differences in their domains: the Jaccard and Aitchison distances, as well as the Kullback-Leibler and the Jensen-Shannon divergences, require vectors with non-negative elements, which means that they can only be safely applied to vectors from Models C and U; and the Kullback-Leibler and Jensen-Shannon divergences, as well as the Aitchison distance, are not defined for item vectors with zero-valued elements. This limitation was overcome by smoothing the item vectors with the Bayesian Multiplicative Treatment with Perks prior when needed (Egozcue et al., 2011). In the config, distances are coded as follows: 0- Euclidean, 1- cosine, 2- Jaccard, 3- Kullback-Leibler, 4- Jensen-Shannon, and 5- Aitchison.

#### 4.1.5 Recommendation Algorithms

Four algorithms were selected, encompassing collaborative filtering or content-based approaches, and also neighbourhood and factorisation techniques. It must be noted that the choices of item representation and comparison adopted by these algorithms move from being independent of those used in surprise assessment (algorithm FP) towards using the same definitions adopted by the surprise metric (algorithm N3):

- *PureSVD (FP)* (Cremonesi et al., 2010): the predicted score,  $\hat{r}_{ui}$ , accounts for the main effects of and interactions between user and item factors, and is exclusively computed from ratings:  $\hat{r}_{ui} = \mathbf{r}_u \cdot \mathbf{Q} \cdot \mathbf{q}_i^T$ , with the matrix  $\mathbf{Q}$  being obtained by factorising the user-item matrix (see Model U). The score  $\hat{r}_{ui}$  is not influenced by the item representation and distance function in the config.
- *Item-kNN with shared distance (N1)* (Koren and Bell, 2015): in a similar vein, the predicted score is exclusively computed from ratings using a neighbourhood approach, and also accounts for main effects and interactions between user and item factors:

$$\hat{r}_{ui} = b_{ui} + \frac{\sum_{j \in S^k(i,u)} s_{ij} (r_{uj} - b_{uj})}{\sum_{j \in S^k(i,u)} s_{ij}}, \text{ with } \quad (13)$$

$$b_{ui} = \mu + b_u + b_i, \text{ and } s_{ij} = \frac{n_{ij} - 1}{n_{ij} - 1 - \lambda} \rho_{ij}.$$

The neighbourhood  $S^k(i, u)$  is the set that contains the  $k$ -nearest neighbours of  $i$  that have been rated by  $u$  ( $k = 50$ ). The predicted score is not influenced by the item representation in the config, but depends on its distance function: the similarity weight  $\rho_{ij}$ , as well as the neighbourhood  $S^k(i, u)$ , are obtained by applying the distance function in

the config to item vectors (invariably) obtained from the  $\mathbf{Q}$  matrix (see Model V).

- *Item-kNN with shared representation (N2)*: the predicted score is obtained by Equation 13. However, now the score is influenced by the item representation in the config, but does not depend on its distance function: the similarity weight  $\rho_{ij}$ , as well as the neighbourhood  $S^k(i, u)$ , are obtained by (invariably) applying the cosine similarity to item vectors encoded as specified in the config. As a result, the predicted score may result from ratings and content data.
- *Item-kNN with shared config (N3)*: the predicted score  $\hat{r}_{ui}$  is computed by Equation 13. Now, the score depends on the item representation and distance function in the config: the similarity weight and the neighbourhood are obtained by applying a distance function to item vectors encoded as specified in the config. Thus, the predicted score may combine ratings and content data.

## 4.2 Results

The process was applied to the 56 configs obtained by combining recommendation algorithms (FP, N1, N2, or N3) with compatible pairs of item representation (C, D, V, or U) and distance function (0 to 5). Figure 2 shows the median and the interquartile range of

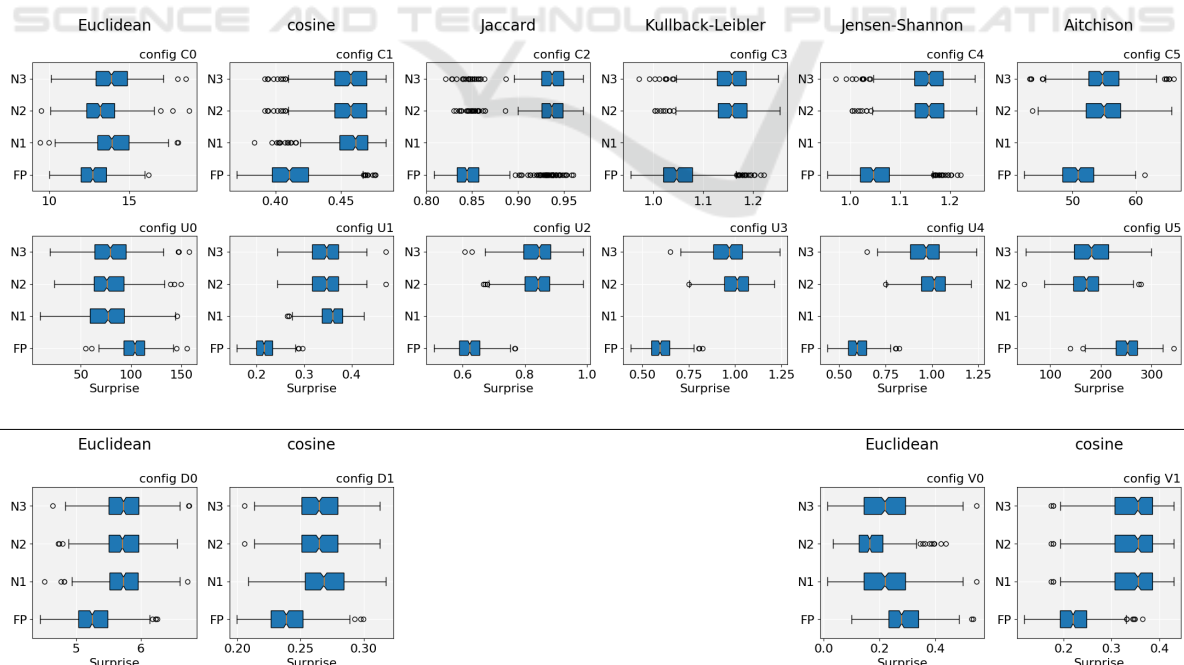


Figure 2: The results obtained from the 56 configs in the experiment: in the first row are the results from Model C; in the second row, from Model U; and from Models D and V in the third row, left and right, respectively. The notches around the median represent its confidence interval. The scale (abscissa) may differ across and within distances.



the surprise obtained from these configs. In summary, these statistics showed that:

1. from the neighbourhood-based algorithms (N1 to N3), the median surprise estimates that were obtained are statistically indiscernible from one another (exceptions in C0, V0, U1, U3, U4, and U5);
2. the median surprise obtained from the factorisation algorithm (FP) is lower than that obtained from the neighbourhood algorithms (exceptions in U0, V0, and U5).

To ensure soundness in the statistical analysis, the configs were arranged into eight groups (Figure 3). In each group, the shaded area delimits a  $2 \times 6$  (Groups 1, 5 and 7) or  $4 \times 2$  (Groups 2, 4, 6 and 8) factorial experiment design. Owing to particular incompatibilities<sup>9</sup> that arise in Group 3, it becomes a subset of the Group 4 and, for that reason, it is not separately considered in the analysis.

To support the discussion of the next findings, consider the pairwise distance distributions shown in Figure 4; they vary in response to the item representation and distance function specified in a config. It must be noted that some distributions are negatively skewed (cosine, Jaccard, Kullback-Leibler, and Jensen-Shannon), and others are positively skewed (Euclidean and Aitchison).

A repeated-measures ANOVA performed on the results from Groups 1, 5, and 7 allowed us to explore variances related to distance functions. At a significance level of  $p < 0.05$ , all main effects and interactions were significant, and contrasts revealed that:

1. the effect of distances with negatively skewed distributions (compared to other forms) in increasing surprise was significant; for the FP algorithm, this increase was larger for Model C ( $r = 0.197$ ), whereas for algorithms N2 and N3 the increase was larger for Model U ( $r = 0.211$  and  $0.264$ , respectively);
2. for negatively skewed distances, no significant difference in surprise between informational (Kullback-Leibler and Jensen-Shannon) and the other distances (cosine and Jaccard) was obtained;
3. for informational distances, no significant difference was obtained between the Kullback-Leibler (asymmetric) and the Jensen-Shannon (symmetric, based on the former).

A similar analysis performed on the results from Groups 2, 4, 6, and 8 allowed us to explore variances

<sup>9</sup>As the N1 algorithm uses item vectors from the  $\mathbf{Q}$  matrix, which has negative elements, some distances can not be safely applied.

related to item representation. At  $p < 0.05$ , all main effects and interactions were significant, and contrasts revealed that:

1. the effect of factorised models (compared to sparse models) in decreasing surprise was significant only for the FP algorithm and was smaller for Euclidean distance than for cosine distance ( $r = 0.223$ );
2. for sparse models (C and U), the effect of using content data (compared to ratings data) in increasing surprise was significant for algorithm FP, and smaller for Euclidean distance than for cosine distance ( $r = 0.252$ );
3. for factorised models (D and V), the effect of using content data (compared to ratings data) in increasing surprise was significant for the neighbourhood-based algorithms N1, N2, and N3 ( $r = 0.492, 0.461$ , and  $0.483$ , respectively), and smaller for cosine than for Euclidean distance.

## 5 DISCUSSION AND CONCLUSION

The aim of this work was to assess the effect that item representation and item comparison models exert on surprise in recommender systems. We started by devising a systematic procedure to (a) identify a set of essential characteristics shared by two competing models of surprise in the literature on cognitive science, (b) find conceptual correlates of these characteristics in the metrics for surprise of recommender systems, and (c) select the surprise metric that is in higher agreement to the cognitive models of surprise. We then applied the selected metric to empirically assess the effects that distinct models of item representation and comparison (i.e. how item vectors are obtained and how the similarity between them is computed), as well as recommendation algorithms, exert on surprise.

Our findings indicate that configs with item comparison models (distance functions in Section 4.1.4) that produce negatively skewed pairwise distributions obtained higher levels of surprise in recommenders that employ sparse representation models (Models C and U in Section 4.1.3). In addition, employing content data in the neighbourhood-based algorithms (N1 to N3 in Section 4.1.5) increased surprise in recommenders with factorised models. It seems that the discriminative power of a distance function, which is reflected in the skew of its pairwise distribution, directly relates to the level of surprise of the system. An implication of this relationship is that the pairwise dis-

Group 1 (FP)						Group 3* (N1)						Group 5 (N2)						Group 7 (N3)					
C0	C1	C2	C3	C4	C5	C0	C1	-	-	-	-	C0	C1	C2	C3	C4	C5	C0	C1	C2	C3	C4	C5
U0	U1	U2	U3	U4	U5	U0	U1	-	-	-	-	U0	U1	U2	U3	U4	U5	U0	U1	U2	U3	U4	U5
D0	D1	-	-	-	-	D0	D1	-	-	-	-	D0	D1	-	-	-	-	D0	D1	-	-	-	-
V0	V1	-	-	-	-	V0	V1	-	-	-	-	V0	V1	-	-	-	-	V0	V1	-	-	-	-

Group 2 (FP)						Group 4 (N1)						Group 6 (N2)						Group 8 (N3)					
C0	C1	C2	C3	C4	C5	C0	C1	-	-	-	-	C0	C1	C2	C3	C4	C5	C0	C1	C2	C3	C4	C5
U0	U1	U2	U3	U4	U5	U0	U1	-	-	-	-	U0	U1	U2	U3	U4	U5	U0	U1	U2	U3	U4	U5
D0	D1	-	-	-	-	D0	D1	-	-	-	-	D0	D1	-	-	-	-	D0	D1	-	-	-	-
V0	V1	-	-	-	-	V0	V1	-	-	-	-	V0	V1	-	-	-	-	V0	V1	-	-	-	-

Figure 3: Eight arrangements of configs. The codes indicate item representations (1<sup>st</sup> character, see *Item representation*) and distance functions (2<sup>nd</sup> character, see *Item comparison*). Incompatible combinations are marked with a “-”.

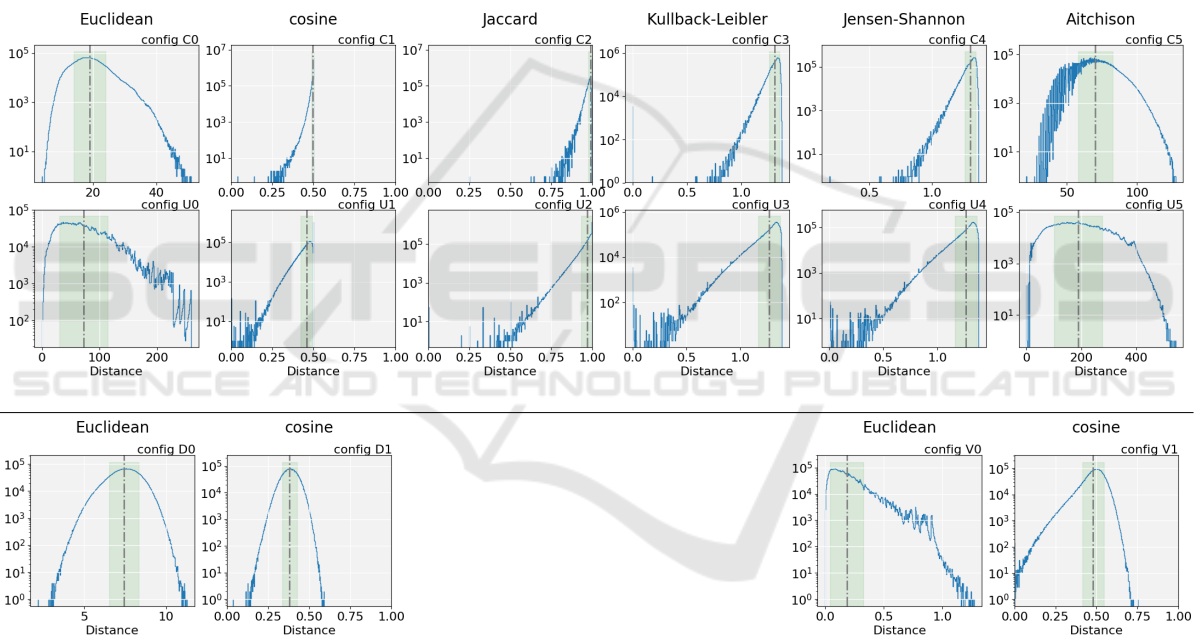


Figure 4: Histograms of pairwise distances for compatible configs (i.e. combinations of an item representation model and an item comparison model). In the first row there are the results from Model C; in the second row, from Model U; and from Models D and V in the third row, left and right, respectively. The distributions were obtained by computing the distance between all pairs of items. The ordinate is in log scale and shows the number of pairs that keep the distance in the abscissa.

tribution can be used to predict the level of surprise of a system. For example, factorisation approaches are generally assumed to achieve higher accuracy and recall when compared to neighbourhood approaches, at the cost of obtaining lower serendipity. However, as the results show, recommenders with factorised models achieved higher surprise than neighbourhood approaches under certain conditions (positively skewed distance in configs U0, V0, and U5). If one accepts the definition of serendipity as being an interaction

between surprise and relevance, then strategies to increase surprise, at an acceptable cost in other properties, may foster serendipity in recommendations.

In summary, this study corroborates the idea that, in offline experiments, the assessment of how much surprise a recommender embeds in its suggestions heavily depends on how the similarity between items is modelled. In other words, it may be the case that the current metrics for surprise are not able to provide good estimates of the performance of the system.

Finally, it should be noted that the contributions presented here could also benefit other research areas that investigate surprise. We hope this work can show that recommender systems might be a fruitful resource in the investigation of surprise in other research areas.

## REFERENCES

- Adamopoulos, P. and Tuzhilin, A. (2011). On unexpectedness in recommender systems: Or how to expect the unexpected. In *Proceedings of the Workshop on Novelty and Diversity in Recommender Systems at the Fifth ACM International Conference on Recommender Systems, DiveRS @ RecSys 2011*, pages 11–18, New York, NY, USA. ACM.
- Akiyama, T., Obara, K., and Tanizaki, M. (2010). Proposal and evaluation of serendipitous recommendation method using general unexpectedness. In *Proceedings of the Workshop on the Practical Use of Recommender Systems, Algorithms and Technologies at the Fourth ACM International Conference on Recommender Systems, PRSAT @ RecSys 2010*, pages 3–10, New York, NY, USA. ACM.
- Baroni, M., Dinu, G., and Kruszewski, G. (2014). Don't count, predict! a systematic comparison of context-counting vs. context-predicting semantic vectors. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 238–247, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Barto, A., Mirolli, M., and Baldassarre, G. (2013). Novelty or surprise? *Frontiers in Psychology*, 4:907.
- Bazerman, M. H. and Moore, D. A. (2009). *Judgment in managerial decision making*. Wiley, Hoboken, NJ, USA, 7th. edition.
- Bellogin, A., Castells, P., and Cantador, I. (2011). Precision-oriented evaluation of recommender systems: An algorithmic comparison. In *Proceedings of the Fifth ACM Conference on Recommender Systems, RecSys '11*, pages 333–336, New York, NY, USA. ACM.
- Bouma, G. (2009). Normalized (pointwise) mutual information in collocation extraction. In *Proceedings of the Conference of the German Society for Computational Linguistics and Language Technology, GSCL 2009*, pages 31–40, Manheim, Germany. GSCL e.V.
- Cremonesi, P., Koren, Y., and Turrin, R. (2010). Performance of recommender algorithms on top-n recommendation tasks. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 39–46. ACM.
- Cremonesi, P., Turrin, R., Lentini, E., and Matteucci, M. (2008). An evaluation methodology for collaborative recommender systems. In *International Conference on Automated Solutions for Cross Media Content and Multi-Channel Distribution, AXMEDIS 2008*, pages 224–231, Washington, DC, USA. IEEE.
- de Gemmis, M., Lops, P., Musto, C., Narducci, F., and Semeraro, G. (2015). Semantics-aware content-based recommender systems. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook*, chapter 26, pages 119–159. Springer, New York, NY, 2nd. edition.
- Deza, M. M. and Deza, E. (2009). *Encyclopedia of Distances*. Springer, Berlin, Heidelberg.
- Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L., and Mateu-Figueras, G. (2011). *Compositional Data Analysis*, chapter 11, pages 139–157. Wiley-Blackwell, Hoboken, NJ, USA.
- Foster, M. I. and Keane, M. T. (2015). Why some surprises are more surprising than others: Surprise as a metacognitive sense of explanatory difficulty. *Cognitive Psychology*, 81:74–116.
- Ge, M., Delgado-Battenfeld, C., and Jannach, D. (2010). Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems, RecSys '10*, pages 257–260, New York, NY, USA. ACM.
- Gershman, S. J. (2017). Similarity as inference. *submitted*.
- Harper, F. M. and Konstan, J. A. (2015). The movielens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 5(4):19.
- Herlocker, J., Konstan, J. A., and Riedl, J. (2002). An empirical analysis of design choices in neighborhood-based collaborative filtering algorithms. *Information retrieval*, 5(4):287–310.
- Herlocker, J. L., Konstan, J. A., Terveen, L. G., and Riedl, J. T. (2004). Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)*, 22(1):5–53.
- Itti, L. and Baldi, P. (2009). Bayesian surprise attracts human attention. *Vision research*, 49(10):1295–1306.
- Kaminskas, M. and Bridge, D. (2014). Measuring surprise in recommender systems. In *Proceedings of the Workshop on Recommender Systems Evaluation: Dimensions and Design, at the 8th ACM Conference on Recommender Systems, REDD @ RecSys '14*, pages 393–394, New York, NY, USA. ACM.
- Kaminskas, M. and Bridge, D. (2016). Diversity, serendipity, novelty, and coverage: A survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Trans. Interact. Intell. Syst.*, 7(1):2:1–2:42.
- Koren, Y. and Bell, R. (2015). Advances in collaborative filtering. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook*, chapter 3, pages 77–118. Springer, New York, NY, 2nd. edition.
- Levy, O. and Goldberg, Y. (2014). Neural word embedding as implicit matrix factorization. In *Advances in neural information processing systems (NIPS)*, pages 2177–2185.
- Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL-02 Workshop on Effective Tools and Methodologies for Teaching Natural Language Processing and Computational Linguistics - Volume 1, ETMTNLP '02*, pages 63–70, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press, New York, NY, USA.

- McNee, S. M., Riedl, J., and Konstan, J. A. (2006). Being accurate is not enough: How accuracy metrics have hurt recommender systems. In *Extended Abstracts on Human Factors in Computing Systems, CHI EA '06*, pages 1097–1101, New York, NY, USA. ACM.
- Meyer, W.-U., Reisenzein, R., and Schützwohl, A. (1997). Toward a process analysis of emotions: The case of surprise. *Motivation and Emotion*, 21(3):251–274.
- Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., Red Hook, NY, USA.
- Mourão, F., Rocha, L., Araújo, C., Meira Jr, W., and Konstan, J. (2017). What surprises does your past have for you? *Information Systems*, 71:137–151.
- Murakami, T., Mori, K., and Orihara, R. (2008). Metrics for evaluating the serendipity of recommendation lists. In Satoh, K., Inokuchi, A., Nagao, K., and Kawamura, T., editors, *New Frontiers in Artificial Intelligence*, pages 40–46, Berlin, Heidelberg. Springer.
- Ning, X., Desrosiers, C., and Karypis, G. (2015). A comprehensive survey of neighborhood-based recommendation methods. In Ricci, F., Rokach, L., and Shapira, B., editors, *Recommender Systems Handbook*, chapter 2, pages 37–76. Springer, New York, NY, 2nd. edition.
- Porter, M. F. (2001). Snowball: A language for stemming algorithms.
- Řehůřek, R. and Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA.
- Reisenzein, R., Horstmann, G., and Schützwohl, A. (2017). The cognitive-evolutionary model of surprise: A review of the evidence. *Topics in Cognitive Science*, (Online Early View).
- Silveira, T., Rocha, L., Mourão, F., and Gonçalves, M. (2017). A framework for unexpectedness evaluation in recommendation. In *Proceedings of the Symposium on Applied Computing, SAC '17*, pages 1662–1667, New York, NY, USA. ACM.
- Turney, P. D. and Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37(1):141–188.
- Tversky, A. (1977). Features of similarity. *Psychological Review*, 84(4):327.
- Tversky, A. and Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131.
- Zhang, Y. C., Séaghdha, D. O., Quercia, D., and Jambor, T. (2012). Auralist: Introducing serendipity into music recommendation. In *Proceedings of the Fifth ACM International Conference on Web Search and Data Mining, WSDM '12*, pages 13–22, New York, NY, USA. ACM.