The WikiWooW Dataset: Harnessing Semantic Similarity and Clickstream-Data for Serendipitous Hyperlinked-Paths Mining in Wikipedia

Cosimo Palma[1][a] and Bence Molnár[2]
[1] *University of Pisa, University of Naples "L'Orientale", Italy*
[2] *University of Pécs, Hungary*

Keywords: Serendipity, Interestingness, Wikipedia, DBpedia, Clickstream-Data, Semantic Similarity, Knowledge Graphs.

Abstract: This paper introduces *WikiWooW*, a dataset generator designed for distilling a formal model of Wikipedia entity-pairs serendipity. The task, foundational to mining serendipitous hyperlinked *paths*, builds upon cognitive theory and exploits serendipity sub-components: graph centrality, popularity, clickstream, corpus-based and knowledge-based similarity. Two proof-of-concept experiments were conducted, based on two different datasets. The first one uses a single Wikipedia entity linked through the DBpedia *dbo:wikiPageWikiLink* property to other 413 entities. These pairs are searched in Wikimedia clickstream data and scored for interestingness according to a principled mathematical model, which is validated against Amazon Mechanical Turk- and author annotations. The second dataset contains 146 random Wikipedia entity-pairs annotated by 10 postgraduate students following detailed guidelines. Average serendipity scores are then correlated with dataset features using the original model and four alternatives. The proposed dataset-generator aims to support Serendipity Mining for Computational Creativity, particularly Knowledge-based Automatic Story Generation, where serendipity matters more than similarity-based interestingness metrics. First results, despite their limitations, confirm the principles initially deduced for modelling serendipity, showing that serendipity can be effectively modeled through comprehensive parameter optimization.

1 INTRODUCTION

From a cognitive perspective, the most salient connections, known as *hard beliefs*, are both easily activated and highly resistant to change.

Their surprisal-level is low.

Recent findings in psychology confirm that surprise is summoned by unexpected (schema-discrepant) events and its intensity is determined by the degree of schema-discrepancy (Reisenzein et al., 2019). Intuitively, humans find interesting not obvious, yet at the same time not random facts.

The typical data structure for representing *facts* is the *Knowledge Graph*, which consists of semantic relationships, i.e. typically *unweighted*, labelled edges between entity nodes (Hogan et al., 2022). Encoding the strength of a link, either in terms of surprisal or according to any other measure, cannot be achieved but by means of numerical values. Enhancing knowledge graphs with *weighted* relationships can enable more nuanced analytical approaches, including centrality measures and community detection algorithms that account for relationship strength (Ristoski and Paulheim, 2016), as well as other network analysis methods such as spectral clustering approaches and information diffusion models that currently find limited application in semantic networks (Bojchevski and Günnemann, 2020).

The implementation of weighting in Linked Open Data (LOD) is attempted in (Hees, 2018) by gamifying data acquisition tasks, thus building the necessary ground truth for validating whether Linked Data can effectively model human associative thinking.

DBpedia-NYD addresses the lack of large-scale benchmarks for assessing the different approaches to the automatic computation of *semantic relatedness* in DBpedia links by providing a synthetic silver standard benchmark with symmetric/asymmetric similarity values from web search data (Paulheim, 2013).

However, despite these contributions, weights based on *serendipity* to enrich DBpedia relationships represents a research direction which still eludes the

[a] https://orcid.org/0000-0002-8161-9782

Palma, C. and Molnár, B.
The WikiWooW Dataset: Harnessing Semantic Similarity and Clickstream-Data for Serendipitous Hyperlinked-Paths Mining in Wikipedia.
DOI: 10.5220/0013747000004000

In *Proceedings of the 17th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2025) - Volume 1: KDIR*, pages 415-425

attention of the scientific community (1.1).

1.1 Related Work

The concept of *interestingness* varies by discipline. In (Hilderman and Hamilton, 1999), a comprehensive survey of measures in Knowledge Discovery (KD)[1] ranks them by *representation* (dataset format: classification rules, summaries, association rules), *foundation* (probabilistic, distance-based, syntactic, utilitarian), *scope* (single rule/rule set) and *class* (objective/-subjective).
Subjective measures involve user's background knowledge (bias, constraints, beliefs, expectations, interactive feedback) and are integrated into the mining process.

Objective measures rely solely on data without user inputs. To this regard, particularly relevant is Silbershatz and Tuzhilin's Interestingness, measuring how *soft beliefs* change with new evidence.

Serendipity is difficult to define and model computationally (Kotkov et al., 2016; McCay-Peet and Toms, 2015).
Generally, serendipity in recommendation systems is calculated as a ratio between *unexpectedness* and *relevance/usefulness*. All serendipity metrics include user preference data, tailored for individuals. For economy constraints, they are not further discussed here. Generally, they stem from sub-component assessment aligning with our paper's components, representing an attempt at modeling serendipity in a principled, *holistic* fashion without individual user subjectivity.

Formal serendipity definition for recommendation systems assumes users have goals (e.g., acquiring items), but *web browsing* is often erratic. For this reason, manual evaluation followed established principles from psychology and Information Retrieval: Silvia's (Silvia, 2009) appraisal theory links interest to novelty, complexity, comprehensibility, driving curiosity and information-seeking and Belkin's (Belkin, 2014) "anomalous states of knowledge" (ASK) views interestingness as information's degree of knowledge-gap resolution. Schmidhuber's (Schmidhuber, 2010) theory suggests interest arises when data balances novelty and comprehensibility, where understanding improves but remains incomplete.

1.2 Problem Statement and Paper's Contribution

A serendipity model beyond similarity-based recommendation systems in databases appears missing in the literature. This problem is in good part due to the Knowledge Graph data-structure of the Web, not originally intended to be weighted, though its expressivity allows such a setting (see section [7.1]). Formulating serendipity for Wikipedia-entities remains open: *interestingness* uses association rules based on subjective user preferences tuned on similarity. Serendipity itself, mainly used in recommendation systems, adds *unexpectedness* to interestingness but remains user-dependent. Leveraging a knowledge graph for calculating serendipity means detaching from the *subjective* dimension of user-based data to distill an objective measure based on axioms grounded in cognitive sciences transformed in logical clauses, in turn mathematically rendered by means of T-Norm conversion[2].

The proposed measure for Serendipity diverges from typical literature definitions. Renouncing subjectivity, it captures curiosity towards the unknown rather than positive response to novel discovery. The resource presented in this article was principally built to investigate correlations between automatically-mined serendipity sub-components and human serendipity scoring. However, as further expounded in the "Future Work" section [7], it can serve a variety of further purposes. The paper's main contribution is an exhaustive method for distilling a serendipity formula for Wikipedia entity pairs, applicable (with due caveats) for building the weighted semantic network foreshadowed in the introduction. Subsequently, retrieving serendipitous *paths* represents a distinct challenge, as cognitive principles determining path interestingness don't fully align with those for simple entity associations (Palma, 2023). This research direction is primary to Computational Creativity, which also emphasizes metrics like *novelty*, *usefulness* and *unexpectedness* (Chhun et al., 2022).

The codebase for dataset generation and analysis is freely accessible at https://github.com/Glottocrisio/WikiWooW[3].

[1] Knowledge Discovery in databases (aka *Knowledge Mining*) identifies previously undiscovered, potentially valuable patterns in extensive databases using diverse techniques and algorithms.

[2] *Objective* is in our final experiment setting equated to *inter-subjective*. *Consensus* on serendipity among users will be used as a heuristic of objectivity, which will be further materialized in the distilled mathematical formula.

[3] It has been implemented in Python 3.9, requiring 2000 lines of code. Processor: Intel(R) Core(TM) i7-6600U CPU @ 2.60 GHz. Run time for the first dataset: 1h47min; second dataset: 2h19min.

2 A NOVEL MODEL OF SERENDIPITY

An entity in a Cross-domain Knowledge Graph like *DBpedia* can be classified by measures simpler to model mathematically than serendipity, such as *popularity*. Intuitively, more page accesses indicate higher popularity. However, if two entities have equal views (clickstream[4]), the one with fewer incoming links should be deemed more popular, since access is less immediate. Among node centrality measures, we selected PageRank centrality (Page et al., 1999) for its off-the-shelf availability via Wikifier API (Brank et al., 2017)[5] and consideration of page *findability* (as for DBpedia Relatedness and Similarity, as well as Cosine Similarity values range from 0 to 1). This heuristic models as following:

$$\mathfrak{P}(i) \simeq \frac{clickstream(i)}{C_{\text{PageRank(i)}}}$$

...where $\mathfrak{P}$ is *Popularity* and *C Centrality*. *Clickstream* and *Popularity* of a node are usually highly correlated. Among *relative measures*, *similarity* is definitely the most known in literature.

Two similarity types exist: *corpus*- and *knowledge*-based (Mihalcea et al., 2006). Capturing this gap delivers the interestingness degree. Knowledge-based similarity uses Relatedness:
$\text{relatedness}(a,b) = \frac{\log(\max(|A|,|B|)) - \log(|A \cap B|)}{\log(|W|) - \log(\min(|A|,|B|))}$
where a and b are the two articles of interest, A and B are the sets of all articles that link to a and b respectively, and W is set of all articles in Wikipedia (Milne and Witten, 2008). Corpus similarity of two entities is calculated by Pointwise Mutual Information or Latent Semantic Analysis of the related concept word lists, based on large corpora (Mihalcea et al., 2006). Having decomposed *serendipity* into modelable components, we proceed toward a general model as in (Palma, 2024)[6]:

$$\mathfrak{I}_{n1,n2} \cong \frac{\mathfrak{P}_{n1,n2} \times \mid \mathfrak{S}_{n1,n2} - \text{ DBpediaRel }_{n1,n2} \mid}{\log_{10}(clickstream_{n1,n2})} \quad (1)$$

where:

$$\mathfrak{S}_{n1,n2} \cong \ln\left(\left|\frac{(\text{CosineSim}_{\text{n1,n2}} + \text{DBpediaSim }_{\text{n1,n2}})}{2}\right|\right)$$

[4]"Clickstream" sometimes synonymous with "click-path" (sequence of hyperlinks visitors follow). Here, it denotes *amount of accesses* between pages (entity pair) or to a single page (single entity).

[5]https://wikifier.org/.

[6]We use "I" instead of the more intuitive "S" for "Serendipity" to differentiate it from the Similarity ("S") measures.

The distributional similarity between two entities is obtained through the simple average between cosine similarity (corpus-based) and DBpedia similarity, whereas the joint popularity has been modelled in order to capture the following set of constraints:

1. The overall serendipity increases if both entities have a high popularity;
2. The overall serendipity increases if one entity is considerably more popular than the other.

The first formulation is "affect-driven" interestingness (Hidi, 2006), while the second is "gap-driven" (Belkin, 2014). The first constraint assumes that popular entities drive attention and curiosity. A high degree of interestingness is also identified in the association of two entities with huge gap in popularity, for two reasons: a gap, as pointed out before, is always interesting; secondly, what is unexpected is also considered interesting. These two conditions have been integrated to model popularity as: $\mathfrak{P}_{n1,n2} \cong \ln(\mathfrak{P}_{n1+n2} + |\mathfrak{P}_{n1-n2}|)$.

3 ALTERNATIVE APPROACHES FOR THE COMPUTATION OF SERENDIPITY IN WikiWooW

Since an entity/node can be either popular or unpopular and a link as corpus- or knowledge-based, we want now to find out how many possible combinations of those elements can occur, taking into account the following constraints:

- Entities can be either popular or unpopular (*absolute measures*);
- The relationship between two entities can be labelled according to both definitions of similarity (*relative measures*).

According to this, our problem can be modelled as a permutation with repetition: if the node can be of two types and the relationship of four types (namely, all possible combinations of two similarities, whose one is knowledge- and the other corpus-based), we can have: $2 \times 4 \times 2 = 16$ possibilities, among which the following five are the only ones showing the previously mentioned gap/contradiction principle:

1. Popular Entity (-) high corpus- AND knowledge-based similarity (-) Unpopular Entity;
2. Popular Entity (-) high corpus- BUT NOT knowledge-based similarity (-) Unpopular Entity;
3. Popular Entity (-) high knowledge- BUT NOT corpus-based similarity (-) Unpopular Entity (e.g. Trivia);

4. Popular Entity (-) high corpus- BUT NOT knowledge-based similarity (-) Popular Entity;
5. Popular Entity (-) high knowledge- BUT NOT corpus-based similarity (-) Popular Entity[7];

Though defining popularity thresholds is inherently fuzzy, we established workable boundaries through trial and error. Based also on annotator consultation, we classified entities with fewer than 6,000 monthly pageviews as unpopular and those exceeding 12,000 as popular.

3.1 Serendipity Models Using Lukasiewicz Operators

The literature presents several approaches to convert logical operators into mathematical ones. To produce the candidate modelings to be tested against the human evaluation, we refer only to the basic operations as listed in *Lyrics* (Marra et al., 2019) (see Figure 1).

op \ t-norm	Product	Lukasiewicz	Gödel
$x \wedge y$	$x \cdot y$	$\max(0, x+y-1)$	$\min(x, y)$
$x \vee y$	$x+y-x \cdot y$	$\min(1, x+y)$	$\max(x, y)$
$\neg x$	$1-x$	$1-x$	$1-x$
$x \Rightarrow y$	$x \leq y ? 1 : \frac{y}{x}$	$\min(1, 1-x+y)$	$x \leq y ? 1 : y$

Figure 1: Logic operation/T-norm conversion table from Lyrics (Marra et al., 2019).

To showcase how to apply the T-norm, in the following we adopt only Lucasiewicz operators.

The used variables are hereby defined:

- $P(e)$: Popularity of entity e (normalized to $[0,1]$);
- $C(e_1, e_2)$: Corpus-based similarity;
- $K(e_1, e_2)$: Knowledge-based similarity;
- $S(e_1, e_2)$: Serendipity of the relationship.

We can express a basic serendipity function as:

$$S(e_1,e_2) = \text{PopularityContrast}(e_1,e_2) \cdot \cdot \text{SimilarityAsymmetry}(e_1,e_2) \quad (2)$$

where:

$$\text{PopularityContrast}(e_1,e_2) = \begin{cases} \min(2, P(e_1)+P(e_2)), \text{if } P(e_1) > \tau_p \wedge P(e_2) > \tau_p \\ \max(0, P(e_1)-P(e_2)), \text{otherwise} \end{cases} \quad (3)$$

and:

$$\text{SimilarityAsymmetry}(e_1,e_2) = \max(0, C(e_1,e_2)+K(e_1,e_2)-1) \quad (4)$$

3.2 Further Alternative Serendipity Models

The following alternative mathematical modelings of serendipity have been computed on the basis of the features extracted through the *WikiWooW* project, with the goal of individuating the one which most resembles the values of the human annotations and evaluations (refer to section [5])[8].

Notation:

- $H(\cdot,\cdot)$: Harmonic mean function;
- $A(\cdot,\cdot)$: Arithmetic mean function;
- R: Resultant length in circular statistics.

Model 2: Logarithmic Popularity Contrast with Similarity Divergence.

$$C^{(2)}_{pop}(e_1,e_2) = \log\left(1 + \frac{\max(P(e_1),P(e_2))}{\min(P(e_1),P(e_2))+1}\right) \quad (5)$$

$$A^{(2)}_{sim}(e_1,e_2) = |S_{cos}(e_1,e_2) - S_{dbp}(e_1,e_2)| + |S_{cos}(e_1,e_2) - R_{dbp}(e_1,e_2)| \quad (6)$$

Model 3: Entropy-Based Popularity with KL Divergence Proxy.

$$C^{(3)}_{pop}(e_1,e_2) = -p_1\log(p_1) - p_2\log(p_2) \quad (7)$$

$$\text{where } p_i = \frac{P(e_i)}{P(e_1)+P(e_2)} \quad (8)$$

$$A^{(3)}_{sim}(e_1,e_2) = \left| s'_{cos} \log\left(\frac{s'_{cos}}{\bar{s}}\right)\right| + \left| s'_{dbp} \log\left(\frac{s'_{dbp}}{\bar{s}}\right)\right| \quad (9)$$

$$s'_i = s_i + 0.1, \quad \bar{s} = \frac{s'_{cos}+s'_{dbp}+s'_{rel}}{3} \quad (10)$$

[7] An example of unexpected conceptual relation between popular entities with already known knowledge-based relationship is the one relating Casanova and Goldoni. It is renown that they were both active in the eighteenth-century Venice, but few know that they are linked through Zanetta Farussi, the mother of Giacomo Casanova, and one of the actresses of Carlo Goldoni. This piece of information can be fully exploited to conceive a story, and would be retrieved by this taxonomy.

[8] The alternative models presented in this subsection have been brainstormed and mathematically rendered with support of Artificial Intelligence. Human intervention encompassed prompting, output analysis, selection, clean-up and correction.

Model 4: Harmonic Mean Contrast with Weighted Variance.

$$C^{(4)}_{pop}(e_1,e_2) = \log\left(\frac{A(P(e_1),P(e_2))}{H(P(e_1),P(e_2))}+1\right) \quad (11)$$

$$A^{(4)}_{sim}(e_1,e_2) = \frac{1}{3}\sum_{i=1}^{3}(s_i-\bar{s})^2\cdot(1+|s_i-\bar{s}|) \quad (12)$$

Model 5: Geometric Dispersion with Circular Variance.

$$C^{(5)}_{pop}(e_1,e_2) = \log(D)\cdot\log(G) \quad (13)$$

$$D = \frac{\max(p_1,p_2)}{\min(p_1,p_2)}, \quad G = \sqrt{p_1\cdot p_2} \quad (14)$$

$$p_1 = P(e_1,e_2)+1, \quad p_2 = |P(e_2,e_1)|+1 \quad (15)$$

$$A^{(5)}_{sim}(e_1,e_2) = 1-R+0.1 \quad (16)$$

$$R = \sqrt{\bar{c}^2+\bar{s}^2}, \quad \theta_i = 2\pi s_i \quad (17)$$

$$\bar{c} = \frac{1}{3}\sum_{i=1}^{3}\cos(\theta_i), \quad \bar{s} = \frac{1}{3}\sum_{i=1}^{3}\sin(\theta_i) \quad (18)$$

4 EXPERIMENT 1: FEATURES IMPORTANCE ASSESSMENT

We chose an a-prioristic serendipity formulation to test correlation between subjectivity and objectivity/intersubjectivity via annotations in this and subsequent experiments. Following the theoretical background, WikiWooW was designed with the features: *Entity1; Entity2; ClickstreamEnt1Ent2; PopularityEnt1; PopularityEnt2 (calculated using PageView and PageRank); PopularityDiff; CosineSimilarityEnt1Ent2; DBpediaSimilarityEnt1Ent2; DBpediaRelatednessEnt1Ent2; InterestingnessEnt1Ent2 (all 5 proposed models); Serendipity Ground Truth Values*.

The *clickstream* data on single Wikipedia-entities is collected by means of *MediaWikiAPI* [9]. On the other hand, to fetch the clickstream-data related to entity couples, we exploit Wikimedia Clickstream Data Dumps [10], as performed in the project *WikiNav* [11]. Similarity measures use *Sematch* (Zhu and Iglesias, 2017)[12]. Initially, "Ground Truth Values" were manual interestingness annotations by Amazon Mechanical Turk (AMT) workers[13], guided by: *Would you be interested in deepening the connection between these Wikipedia entities? Does this connection spark your curiosity?* Of 413 entity pairs, all except two were labeled interesting, creating drastic dataset imbalance. To alleviate this, we randomly selected 1000 Wikipedia entity pairs[14], shuffling the original 413 among them. The second annotation formulation omits to make explicit that the entity pairs are linked. This raised pairs to 31 (19 original), encouraging for larger-scale analysis but insufficient for evaluation: individuals with different interests and "interestingness" conceptions unlikely agree so extensively.

For this reason, and given the exploratory research nature, serendipity values were balanced using median threshold for "non-interesting" values.

The related literature proposes, among others, Principal Component Analysis (PCA) and SHapley Additive exPlanations (SHAP) for features importance assessment (FIA).

Principal Component Analysis (PCA) (Jolliffe and Cadima, 2016) identifies key patterns by projecting the original data onto a new set of axes, known as principal components. These components are hierarchically arranged based on their ability to capture data variability, with the first component accounting for the most variance. The unsupervised nature and variance maximization objective make it unsuitable for identifying features most relevant to prediction tasks and should be reserved only for dimensionality reduction.

SHAP (SHapley Additive exPlanations) (Lundberg and Lee, 2017) draws inspiration from cooperative game theory to quantify the contribution of each feature to a model's output.

However, given the characteristics of our dataset and the general superior performance compared to the above-mentioned alternatives for the task of FIA, we have selected the *cforest* function (Strobl et al., 2007), which provides unbiased variable selection within classification trees. When implemented with subsampling without replacement, it produces reliable importance measures robust across predictor variables with different measurement scales or category numbers.

The calculation of feature importance showcased in Table 1 shows how our first attempt of serendipity modeling (ref. to [1]) performs slightly better than all other features, demonstrating that unifying the subcomponents in a single expression might lead to a satisfactory formulation.

[9] https://pypi.org/project/mediawikiapi/.

[10] https://dumps.wikimedia.org/other/clickstream/readme.html. Dataset uses English entities accessed November 2018.

[11] https://wikinav.toolforge.org/.

[12] https://pypi.org/project/sematch/.

[13] Three workers label each entity pair; final value averages their confidence ratings.

[14] Still directly linked through *dbo:wikiPageWikiLink*.

Table 1: Random Forest - Features Importances.

Feature	Importance
Serendipity Model [1]	**0.231**
Clickstream	0.178
DBpediaSimilarity	0.127
DBpediaRelatedness	0.121
CosineSimilarity	0.118
PopularityDiff	0.112

In the following section, we apply another approach to bind the features in an equation which might at best express the serendipity how emerging from the human annotation.

4.1 Symbolic Regression Analysis

Discovering interpretable mathematical expressions directly from data is a task that has traditionally been managed using genetic programming (Vanneschi and Poli, 2012). Recently, however, there has been an increasing interest in employing a deep learning approach for this purpose (Makke and Chawla, 2023). The following expression has been found by using the Python library *gplearn*, and tries to capture the weighting of features to result in the ground truth. Results from other libraries, showing similarly convoluted results, have been omitted for reasons of space.

$$
\begin{aligned}
I(s_1,s_2) = \frac{-0.166}{X_3} - \Bigg(& (0.729 - X_3 + X_5) + \\
& + \Bigg(\frac{0.729 - X_3 + \frac{0.729 - \frac{X_3}{X_4 clickstream_{n1,n2} X_3} + X_5}{X_4}}{0.811} \\
& + \frac{X_5}{X_3} \Bigg) + \left(\frac{\frac{0.647}{X_3} + X_5}{X_4} + X_5 \right) \Bigg) \\
& + \left(\frac{X_5}{X_3} - 0.606 \times clickstream_{n1,n2} \right) + \\
& + \frac{2X_5 - X_3}{X_4} + \frac{X_5}{X_3}
\end{aligned}
$$

where:

$$
\begin{aligned}
X_3 &= \text{CosineSim}_{\text{n1,n2}} \\
X_4 &= \text{DBpediaSim}_{\text{n1,n2}} \\
X_5 &= \text{DBpediaRel}_{n1,n2}
\end{aligned}
$$

Despite hyperparameter optimization efforts, the model's performance remained unchanged, yielding outputs that exhibited either excessive simplicity or unnecessary complexity (as observed in the output proposed above).

5 EXPERIMENT 2: MORE ANNOTATORS, BETTER GUIDELINES

In a last attempt to reconcile subjective evaluations in a seeming inter-subjectivity, a more thorough evaluation has been designed, featuring the drafting of guidelines and the selection of reliable evaluators [15].

Low inter-annotators agreement on serendipity performed on a 60 entity-pairs dataset, representing 10% of the final dataset envisioned in case of successful evaluation has lead to the creation of a smaller dataset of 146 entity pairs, extracted from the Wikimedia Clickstream Data Dump, where the final serendipity value for each entity couple was calculated as an average of all scores.

Table 2: Overall Inter-Annotator Agreement Metrics.

Metric	Value	Interpretation
Fleiss' Kappa	0.38	Fair agreement
Cronbach's Alpha	0.14	Poor reliability
Krippendorff's Alpha	0.09	Slight agreement
Avg. MSE	0.35	–
Avg. Percent Agreement	0.43	–

In order to investigate the link between entity-pair popularity and serendipity, one third of the dataset is comprised of very popular entities (above 12000 page-views), one third of unpopular entities (below 6000), and one third of mixed popularity (one entity popular and the other not).

The following guidelines instruct annotators on assessing entity pairs from *English Wikipedia* for serendipity, defined as: "the property of an entity pair whose relationship is simultaneously **unexpected** and **relevant** or **interesting**." We define "close" relationships as *non-trivial* direct connections between entities. *Trivial* relationships include generic connections like "related to," "same as," "are things," or "are persons." Non-trivial connections include "child of," "successor of," "grown in," "coeval with," and other specific, meaningful relationships. Annotators evaluate entity pairs (e.g., *'Mark Antony'; 'Alexander the Great'*) using a *Google Form* with yes/no questions, resulting in scores of 1 (serendipitous), 0.5 (unexpected but irrelevant), or 0 (too obvious).
Example 1: *'Classical antiquity'; 'Alexander the Great'* Score: **0** (too obvious)

[15]The ten evaluators are Master of Education students from the Faculty of Humanities and Social Sciences at the University of Pécs, Hungary, each one majoring in English language and culture and selected based on their level of English proficiency.

- Q1: Could there be a close relationship? → Yes
- Q2: Do you know the relationship's nature? → Yes (general knowledge)
- Result: Too obvious

Example 2: *'Chanakya (TV series)'; 'Alexander the Great'*.
Score: **0.5** or **1**.
A relationship seems plausible but its nature is unknown. The score depends on curiosity: lack of interest yields 0.5 (irrelevant), while genuine curiosity yields 1 (serendipitous).
Example 3: *'Mark Antony'; 'Alexander the Great'*.
Score: **1** (serendipitous).
Despite different historical periods, a connection appears likely but remains unclear, generating curiosity.
The annotation process captures perceived relatedness (Q1), subjective knowledge (Q2), and serendipity scores, enabling analysis of how perceived connections and prior knowledge influence serendipity judgments and providing empirical data for model refinement.

6 VISUALIZATION AND EVALUATION OF RESULTS

To investigate serendipity behavior across popularity combinations (as shown in section 2), we categorized entity couples into three groups: popular-popular, unpopular-unpopular, and popular-unpopular pairs.

Table 3: Average Serendipity and Knowledge Metrics by Entity Popularity.

Category	PR	SK	Ser.
Overall	0.60	0.40	0.50
Popular	0.59	0.44	0.57
Unpopular	0.60	0.39	0.39
Pop-Unpop	0.62	0.38	0.49

PR: Perceived Relatedness; SK: Subjective Knowledge
Serendipity: 0.6 = Moderate; 0.7 = High; 0.8+ = Very High

Table 3 reveals that popular entities exhibit significantly higher average serendipity (0.572) compared to unpopular entities (0.386), validating our principled assumption from section 2. Furthermore, results confirm that high clickstream correlates with decreased serendipity scores: all highly serendipitous pairs (scores higher than 0.8) consistently associate with low clickstream values. This observation, aligning with our initial axioms, suggests potential optimization strategies for computationally expensive serendipitous couple retrieval algorithms.

The serendipity values' considerable variance around the mean (0.5) validates our choice to calculate serendipity as an average. However, this variance simultaneously demonstrates that objective serendipity measurement (at least within our problem formulation) remains infeasible, at least for small datasets, as the ones adopted for the experiments.
Although deeper analysis is required to identify the best-performing model against ground truth, any resulting model will be far from definitive, having been derived through induction rather than the desirable deductive approach. Nevertheless, these results provide foundational value for serendipity computation in other computational creativity settings, as they emerge from logical, explicit principles readily adaptable to diverse applications.

Figure 3 reveals key differences between popular and mixed entity couples: in the first case, perceived relatedness highly correlates with serendipity, while in the second case only mildly. Interestingly, popular pairs show also a mild correlation with subjective knowledge and clickstream, which conversely anti-correlate for pairs of mixed popularity, where the best predictors are identified with *DBpedia Relatedness* and the mathematical modelling conceptualized in 1. These are amongst the worst predictors in popular couples, where the relevance of clickstream for serendipity annotation also reflects in the alternative serendipity modelling that were tuned with it, performing better than the same without clickstream. For the same reason, an opposite behavior is expected for couples of mixed popularity.

We also notice from the graphics how high serendipity is indeed correlated with considerable differentials between entity popularity values, as well as between Cosine Similarity (together with DBpedia Relatedness) and DBpedia Similarity, observation already postulated in the principled approach.

Figure 4 illustrates model performance across serendipity ranges: all models perform adequately for low serendipity values; however, only Model 2 demonstrates encouraging results for high serendipity cases. Medium serendipity values prove entirely unpredictable across all models, suggesting inherent complexity in this range.

7 FUTURE WORK

Several avenues for future research emerge from this study. First, multivariate analysis can be employed to examine relationships between multiple variables simultaneously. The model itself can even be enhanced by incorporating additional features into the equation, while expanding the number of testers will im-

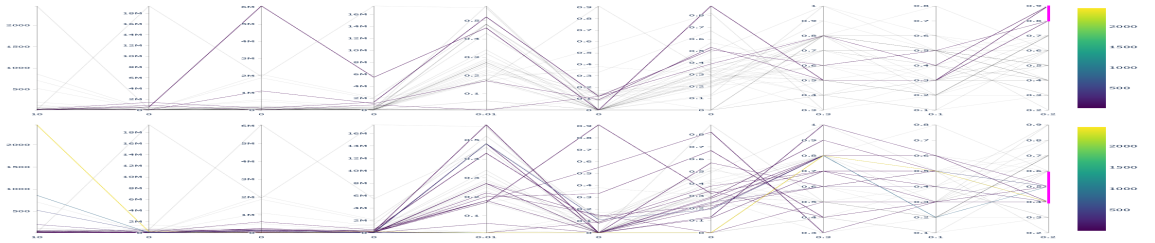

Figure 2: Parallel coordinate plot of serendipity and the extracted features together with the human annotated subcomponents. From left to right: Clickstream, PopularityEnt1, PopularityEnt2, PopularityDiff, Cosine Similarity, DBpedia Similarity, DBpedia Relatedness, Perceived Relatedness, Subjective Knowledge, Serendipity.

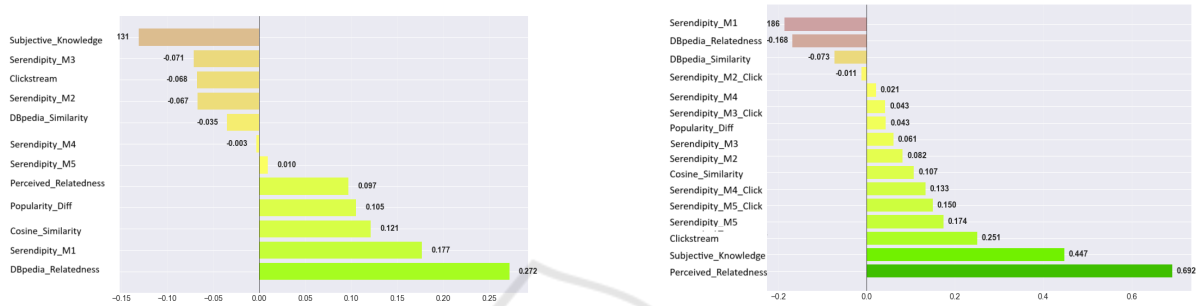

Figure 3: Serendipity predictors based on Pearson correlation. On the left, values computed on entity couples with mixed popularity. The AI-generated alternative serendipity models have not been tuned with clickstream data. On the right, values computed on entity couples with high popularity; clickstream (*Click) included in serendipity models.

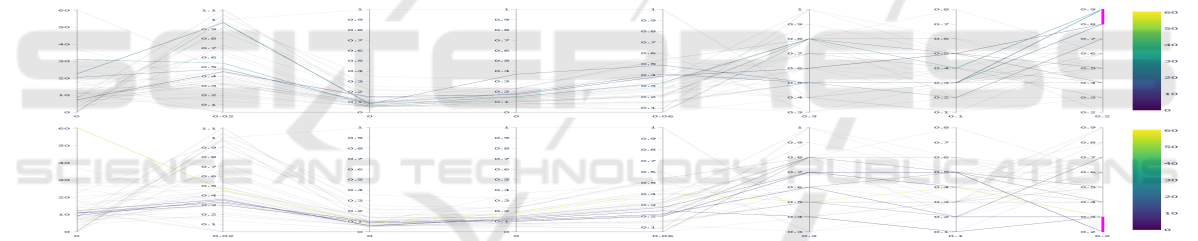

Figure 4: Parallel coordinate plot of serendipity and the alternative models. From left to right: the five serendipity models, Perceived Relatedness, Subjective Knowledge, Serendipity.

prove annotation reliability. The extensive multilingual coverage of constantly-maintained clickstream data enables future cross-cultural studies assessing interestingness variance across languages—with significant implications for Computational Creativity. Additionally, diachronic analysis comparing data-dumps across years would reveal user interest shifts over time.

7.1 Implementing Weighting of Knowledge Graph Relations Through SW-Technologies

To transform traditional knowledge graphs into weighted networks, numerical values can be incorporated directly into the graph structure using Semantic Web (SW) technologies such as RDF-star (RDF*) (Hartig et al., 2021), which enables the annotation of RDF triples with additional metadata.

For instance, the DBpedia property `dbo:wikiPageWikiLink` could be enhanced with narrative interestingness scores by creating qualified statements that attach numerical weights to each link. Using RDF-star syntax, a weighted link could be expressed as,

```
«:Entity1 dbo:wikiPageWikiLink :Entity2»
:weight 0.8
```

enabling SPARQL queries to filter and rank relationships based on their numerical significance:

```
SELECT ?s ?p ?o WHERE {
BIND(<<?s ?p ?o>> AS ?t)
  ?t serendipity ?s .
  FILTER ( ?s > 0.7 )
}
```

Achieving the data coverage necessary for a large knowledge base such as Wikipedia requires a collec-

tive effort, which can only thrive on enhanced User Experience and even gamification strategies.

7.2 Of Colors and Gadgets

Links are fundamental to Wikipedia and web development, though (Dimitrov et al., 2017) found that only about 4% of Wikipedia links receive more than 10 monthly clicks. To understand WikiWooW's potential applications and benefits, we must first examine the various link types, colors, and the concept of "orphan articles" on Wikipedia.

Wikipedia employs three distinct link categories.[16] Internal links (wikilinks) connect pages within the same project, while *interwiki* links connect to different Wikimedia projects using prefixes (e.g., "de" for German Wikipedia). External links, marked with an icon, can direct to any web page. This article focuses exclusively on internal links, which are the sole links considered in our dataset.

Related to link connectivity is the issue of "orphan articles"—Wikipedia pages lacking incoming links from other main namespace pages.[17] Although these pages remain searchable via internal search and external services, Wikipedia's principle advocates for their integration through related page links. Such integration not only increases readership but also attracts contributors who can enhance content quality.

To facilitate navigation and user experience, Wikipedia implements a color-coding system for links.[18] Blue indicates unvisited existing pages, purple shows visited existing pages, red marks non-existent unvisited pages, and light maroon denotes non-existent visited pages. While these colors vary by skin and can be customized through user scripts and CSS, the fundamental principle persists: blue for existing articles, red for non-existent ones.

Building on this customization capability, user scripts commonly modify default link colors, with community-approved scripts called "gadgets" being widely adopted. For instance, the "DisambiguationLinks" gadget highlights disambiguation pages in different colors.[19] Following this model, WikiWooW's values could similarly identify and color-code interesting links based on their clickstream data, graph- or cosine-similarity, or even serendipity. This approach would harness users with a more objective discovery experience while assisting editors in finding less popular articles requiring improvement.

[16]https://en.wikipedia.org/wiki/Help:Link.

[17]https://en.wikipedia.org/wiki/Wikipedia:Orphan.

[18]https://en.wikipedia.org/wiki/Help:Link_color.

[19]https://en.wikipedia.org/wiki/MediaWiki:Gadget-DisambiguationLinks.css.

However, the effectiveness of such color-coding must consider user behavior patterns. (Dimitrov et al., 2016) discovered that readers primarily click links in prominent locations: lead sections, right sidebars with infoboxes, and left body areas, while generally avoiding right-side regions. Consequently, WikiWooW's model could counteract this spatial bias by identifying and marking valuable links in underutilized regions, though initial clickstream data would inevitably reflect these existing patterns.

Ultimately, implementing a user script with WikiWooW's model would enable testing whether users find the suggested links engaging within Wikipedia's native environment: even if an explicit feedback API is not provided, the impact of the proposed enhanced user experience can be implicitly assessed through the shift of clickstreams between entities as they are already monthly collected in the Wikimedia Clickstream Data Dump.

8 LIMITATIONS, CHALLENGES AND FINAL REMARKS

If attempts to modelling serendipity in a principled fashion continue to prove unsatisfactory, a machine learning approach represents the logical next step for modeling the complex interactions among the established features, though careful attention must be paid to avoiding overfitting.

Since harvested values rely primarily on *Sematch*, an application developed nearly a decade ago despite ongoing maintenance, a thorough assessment of value retrieval methods is necessary to purge outdated information from the dataset. Corpus-based similarity could be better captured using word embeddings from *Wikipedia2vec*, which temporally aligns with Wikimedia clickstream data. To enhance precision, clickstream data should be computed as monthly averages rather than our current single-month snapshot (November 2018).

Furthermore, its normalization could explore alternatives that, better than logarithms, can better capture the present remarkable fluctuations. For clickstream-agnostic Knowledge Graph analysis, the equation should gradually de-emphasize clickstream in favor of graph centrality. Adding centrality as an explicit dataset feature (currently implicit in popularity, see equation 2) represents the immediate next step for observing its behavior against annotations. With improved dataset quality and size following these guidelines, even symbolic regression should yield better results.

Figure 3 already demonstrates how clickstream

conserve its relevance in predicting serendipity only for popular-popular entity pairs. Exploring clickstream-centrality correlations could yield satisfactory popularity approximations using centrality alone, enabling analysis on any graph without clickstream requirements.

We have shown how all basic initial assumptions, based on intuition and consolidated from cognitive- and information theory, were validated from the experimental data. Despite a model of Serendipity which clearly outperforms the other has not been found yet, our experimental results show that combination of the individuated sub-components is a proxy for serendipity measure. Beyond these specific directions, the comprehensive nature of our dataset positions it as a valuable resource for broader research in computational creativity, offering multiple possibilities for exploration that constitute a significant contribution in itself.

AUTHORS CONTRIBUTION

Cosimo Palma: Conceptualization, Methodology, Codebase development, Formal Analysis, Investigation, Data Curation, Writing (paper original draft, review and editing, annotators guidelines original draft, review and editing), Information Visualization, Project administration.
Bence Molnár: Methodology, Investigation, Recruitment, Coaching and Supervision of annotators, Writing (section 7.2 original draft, paper review and editing, Annotators guidelines review and editing), Project administration.

ACKNOWLEDGEMENTS

This work could not have been carried out without the support of several scholars. First of all, Dr. Maria Pia Di Buono, who actively participated throughout the ideation and design of the second experiment; her suggestions for the annotator guidelines were particularly instrumental in improving the quality of data collection. The work has also benefited from insightful discussions with Dr. Emanuele Marconato, Philipp Bous, Prof. Dr. Carlo Strapparava, Sebastien Albouze, Dr. Vassilis Tzouvaras, and Dr. Victor De Boer. We extend our sincere gratitude to the data annotators and to the anonymous reviewers, whose constructive feedback significantly contributed to enhancing this paper.

REFERENCES

Belkin, N. (2014). Anomalous states of knowledge as a basis for information retrieval. *Canadian Journal of Information Science*, pages 133–143.

Bojchevski, A. and Günnemann, S. (2020). Adversarial attacks on node embeddings via graph poisoning. In *International Conference on Machine Learning*, pages 695–704. PMLR.

Bordes, A., Usunier, N., Garcia-Duran, A., Weston, J., and Yakhnenko, O. (2013). Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems*, pages 2787–2795.

Brank, J., Leban, G., and Grobelnik, M. (2017). Annotating documents with relevant wikipedia concepts. In *Proceedings of the Slovenian Conference on Data Mining and Data Warehouses (SiKDD 2017)*.

Cheng, G., Gunaratna, K., Thalhammer, A., Paulheim, H., Voigt, M., and García, R. (2015). Sematch: Semantic entity search from knowledge graph. In Cheng, G., Gunaratna, K., Thalhammer, A., Paulheim, H., Voigt, M., and García, R., editors, *Joint Proceedings of the 1st International Workshop on Summarizing and Presenting Entities and Ontologies and the 3rd International Workshop on Human Semantic Web Interfaces (SumPre 2015, HSWI 2015)*, Portoroz, Slovenia.

Chhun, C., Colombo, P., Suchanek, F. M., and Clavel, C. (2022). Of human criteria and automatic metrics: A benchmark of the evaluation of story generation.

Diedrich, J., Benedek, M., Jauk, E., and Neubauer, A. (2015). Are creative ideas novel and useful? *Psychology of Aesthetics, Creativity, and the Arts*, 9:35–40.

Dimitrov, D., Singer, P., Lemmerich, F., and Strohmaier, M. (2016). Visual positions of links and clicks on wikipedia. In *Proceedings of the 25th International Conference Companion on World Wide Web*, WWW '16 Companion, page 27–28, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Dimitrov, D., Singer, P., Lemmerich, F., and Strohmaier, M. (2017). What makes a link successful on wikipedia? In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 917–926, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Dong, X., Gabrilovich, E., Heitz, G., Horn, W., Lao, N., Murphy, K., Strohmann, T., Sun, S., and Zhang, W. (2014). Knowledge vault: A web-scale approach to probabilistic knowledge fusion. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 601–610. ACM.

Freitas, A. A. (1998). On objective measures of rule surprisingness. In Carbonell, J. G., Siekmann, J., Goos, G., Hartmanis, J., Van Leeuwen, J., Żytkow, J. M., and Quafafou, M., editors, *Principles of Data Mining and Knowledge Discovery*, volume 1510, pages 1–9. Springer Berlin Heidelberg, Berlin, Heidelberg. Series Title: Lecture Notes in Computer Science.

Guo, Q., Zhuang, F., Qin, C., Zhu, H., Xie, X., Xiong, H., and He, Q. (2020). A survey on knowledge graph-

based recommender systems. volume 34, pages 3549–3568. IEEE.

Hartig, O., Champin, P.-A., Kellogg, G., and Seaborne, A. (2021). RDF-star and SPARQL-star. W3c community group final report, W3C.

Hees, J. (2018). *Simulating Human Associations with Linked Data*. doctoralthesis, Technische Universität Kaiserslautern.

Hidi, S. (2006). Interest: A unique motivational variable. *Educational Research Review*, 1(2):69–82.

Hilderman, R. J. and Hamilton, H. J. (1999). Knowledge Discovery and Interestingness Measures: A Survey. *Computer Science*, page 28.

Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G. (2013). Yago2: A spatially and temporally enhanced knowledge base from wikipedia. *Artificial Intelligence*, 194:28–61.

Hogan, A., Blomqvist, E., Cochez, M., d'Amato, C., de Melo, G., Gutierrez, C., Gayo, J. E. L., Kirrane, S., Neumaier, S., Polleres, A., Navigli, R., Ngomo, A.-C. N., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S., and Zimmermann, A. (2022). Knowledge Graphs. *ACM Comput. Surv.*, 54(4):1–37. arXiv:2003.02320 [cs].

Itti, L. and Baldi, P. (2006). Bayesian surprise attracts human attention. In *Advances in neural information processing systems*, pages 547–554.

Jolliffe, I. T. and Cadima, J. (2016). Principal component analysis: a review and recent developments. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 374(2065).

Kotkov, D., Wang, S., and Veijalainen, J. (2016). A survey of serendipity in recommender systems. *Knowledge-Based Systems*, 111.

Lundberg, S. M. and Lee, S.-I. (2017). A unified approach to interpreting model predictions. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Makke, N. and Chawla, S. (2023). Interpretable scientific discovery with symbolic regression: A review.

Marra, G., Giannini, F., Diligenti, M., and Gori, M. (2019). Lyrics: a general interface layer to integrate logic inference and deep learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 283–298. Springer.

McCay-Peet, L. and Toms, E. G. (2015). Investigating serendipity: How it unfolds and what may influence it. *Journal of the Association for Information Science and Technology*, 66(7):1463–1476.

Mcgarry, K. (2005). Mcgarry, k.: A survey of interestingness measures for knowledge discovery. know. eng. rev. 20(01), 39-61. *Knowledge Eng. Review*, 20:39–61.

Mihalcea, R., Corley, C., and Strapparava, C. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *AAAI*, volume 6.

Milne, D. and Witten, I. H. (2008). Learning to link with wikipedia. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*. ACM.

Nickel, M., Murphy, K., Tresp, V., and Gabrilovich, E. (2016). A review of relational machine learning for knowledge graphs. *Proceedings of the IEEE*, 104(1):11–33.

Page, L., Brin, S., Motwani, R., and Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab. Previous number = SIDL-WP-1999-0120.

Palma, C. (2023). Modelling interestingness: Stories as L-Systems and Magic Squares. In *Text2Story@ECIR*, Dublin (Republic of Ireland).

Palma, C. (2024). Modelling interestingness: a workflow for surprisal-based knowledge mining in narrative semantic networks. In *SEMMES'24: Semantic Methods for Events and Stories, co-located with the 21th Extended Semantic Web Conference (ESWC2024)*.

Paulheim, H. (2013). Dbpedianyd - a silver standard benchmark dataset for semantic relatedness in dbpedia. In *NLP-DBPEDIA@ISWC*.

Reisenzein, R., Horstmann, G., and Schützwohl, A. (2019). The Cognitive-Evolutionary Model of Surprise: A Review of the Evidence. *Topics in Cognitive Science*, 11(1):50–74.

Ristoski, P. and Paulheim, H. (2016). Semantic web in data mining and knowledge discovery: A comprehensive survey. *Journal of Web Semantics*, 36:1–22.

Schmidhuber, J. (2010). Formal theory of creativity, fun, and intrinsic motivation (1990–2010). *IEEE Transactions on Autonomous Mental Development*, 2:230–247.

Silvia, P. J. (2009). Looking past pleasure: Anger, confusion, disgust, pride, surprise, and other unusual aesthetic emotions. *Psychology of Aesthetics, Creativity, and the Arts*, 3(1):48–51.

Strobl, C., Boulesteix, A.-L., Zeileis, A., and Hothorn, T. (2007). Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformatics*, 8(1).

Trouillon, T., Welbl, J., Riedel, S., Gaussier, É., and Bouchard, G. (2016). Complex embeddings for simple link prediction. In *International Conference on Machine Learning*, pages 2071–2080.

Vanneschi, L. and Poli, R. (2012). Genetic programming — introduction, applications, theory and open issues. In Rozenberg, G., Bäck, T., and Kok, J. N., editors, *Handbook of Natural Computing*. Springer, Berlin, Heidelberg.

Zhu, G. and Iglesias, C. A. (2017). Computing semantic similarity of concepts in knowledge graphs. *IEEE Transactions on Knowledge and Data Engineering*, 29(1):72–85.