A Serendipity Model For News Recommendation

M. Jenders, T. Lindhauer, G. Kasneci, R. Krestel, and F. Naumann

Hasso Plattner Institute
Prof.-Dr.-Helmert-Str. 2-3, 14482 Potsdam, Germany

Abstract. Recommendation algorithms typically work by suggesting items that are similar to the ones that a user likes, or items that similar users like. We propose a content-based recommendation technique with the focus on serendipity of news recommendations. Serendipitous recommendations have the characteristic of being unexpected and yet fortunate and interesting to the user, and thus might yield higher user satisfaction. In our work, we explore the concept of serendipity in the area of news articles in a user study and propose a general framework that incorporates the benefits of serendipity- and similarity-based recommendation techniques and is evaluated against other baseline recommendation models in a final user study.

1 Introduction

Popular recommendation algorithms employ similarity measures to generate their recommendation lists. Their abilities in predicting users' interest can be quantified using *accuracy*- and *relevance*-based measures. They can operate purely on item features or incorporate user profiles with preferences and ratings. While the resulting recommendations are often accurate, they tend to favor popular items or ones that users already know and therefore often miss opportunities to surprise users with items that are to some extent unrelated and unfamiliar, yet satisfactory.

In this paper, we aim to address this issue by focusing on serendipitous recommendations in the area of news articles. Serendipity refers to the event of stumbling upon something that is unexpected and yet useful. For example, when the task is to recommend news articles about "Turkey's EU membership", a traditional recommender would favor articles focusing on this very issue, which, while relevant, do not expand a reader's horizon beyond this topic. Meanwhile, a serendipitous recommendation might contain an article arguing that MTV Turkey already establishes stronger ties between Turkey and the West than an acceptance into the EU ever would. In fact, this notion of serendipity and its usefulness was confirmed in our user evaluation (Section 3).

Indeed, in cases such as the above, unexpected articles may complement similar ones, even if they do not show very high similarity (as an estimation of relevance) to known items. Algorithms that focus on both aspects (i.e., unexpectedness and similarity) can produce more useful recommendation lists by suggesting items that expand the user's horizon in addition to familiar items [7].

Our approach does not assume user profiles to generate recommendations, as this would be a typical case in practice. Since serendipity is subject to general perception, we claim that serendipitous recommendations can be generated from content features alone with sufficient quality.

In an exploratory user study, we evaluated the impact different features (e.g., named entities, the relationship between entities or between latent topics) on serendipity that is perceived by humans and found that *unexpected combinations of latent topics* induce a strong and consistent signal upon which serendipity can be formalized. We used this insight to develop a purely content-based unexpectedness model that is combined with a similarity model for pre-filtering (Section 2) and evaluated it against popular algorithms in a concluding user study (Section 3). All experiments were conducted on the subset of all news articles of the New York Times Corpus[1] published between 2005 and 2007.

We discovered two different facets of serendipity: On one hand, recommendations that display a high similarity to the original article (and could therefore be seen as discussing related and relevant subjects) were judged as surprising and interesting and could therefore be labeled serendipitous. On the other hand, the recommendations of our unexpectedness model were rated as decidedly less relevant with respect to the original article, nevertheless also highly surprising and interesting. In the above example, the aforementioned article about Turkey's MTV station exhibits a strong topic shift and is very unexpected. For the goal of recommending serendipitous articles, our unexpectedness model therefore is not a direct competitor to similarity-based recommendation, but focuses on a different facet of serendipity. We hence combined both models to yield the final serendipity-based recommendation algorithm that presents the best recommendations from both algorithms to the user.

2 A Serendipity Model

In this section, we discuss the creation of a recommendation model for newspaper articles based on serendipity, i.e., on unexpectedness and interestingness. Since the latter is highly user-dependent, we first focused on capturing the general unexpectedness of a document without setting the article in relation to specific other articles. Since the resulting model (Equation 3) does not compare documents, it can only be used to rank a corpus regarding *topical unexpectedness*. We then measured the interestingness of documents through the similarity of the

[1] https://catalog.ldc.upenn.edu/LDC2008T19

suggested article to the article currently being read, as it is the only available indication of user interest.

Our final serendipity model uses a non-linear combination of both models and is described in detail in Section 3. It ranks recommendations separately to each model and then selects the most promising recommendations (for a given article) based on a boosting algorithm.

2.1 Deriving an Unexpectedness Model

Bache et al. developed a model to quantify *document diversity* in the context of scientific papers [3]. The proposed model is given in Equation 1 and estimates the proportion of a topic z_i as its probability according to a latent Dirichlet allocation (LDA) model. $\delta(z_i, z_j)$ is the dissimilarity of two topics z_i, z_j and is estimated based on their co-occurrences across documents in the corpus. $\boldsymbol{d}$ denotes a document as a vector of term frequencies, while $div(\boldsymbol{d})$ expresses the topic diversity in a document.

$$div(\boldsymbol{d}) = \sum_{i=1}^{k} \sum_{j=1}^{k} p(z_i|\boldsymbol{d}) \cdot p(z_j|\boldsymbol{d}) \cdot \delta(z_i, z_j) \tag{1}$$

As the model addresses the dissimilarity of topic pairs while considering their proportions in the documents, we base our model on this approach. We limit our model's calculations to the document's main topics $Z_{Main}(\boldsymbol{d})$ and use it to rank documents by their estimated *topical unexpectedness*:

$$u(\boldsymbol{d}) = \sum_{\substack{z_i, z_j \in \\ Z_{Main}(\boldsymbol{d})}} sp(z_i, \boldsymbol{d}) \cdot sp(z_j, \boldsymbol{d}) \cdot dis(z_i, z_j) \cdot c(z_i, z_j, \boldsymbol{d}) \tag{2}$$

$$u_{norm}(\boldsymbol{d}) = u(\boldsymbol{d}) \cdot norm(\boldsymbol{d}, Z_{Main}(\boldsymbol{d})) \tag{3}$$

The model comprises four main components that are constructed with an information-theoretic background, which are detailed in Sections 2.2 to 2.5. The effects of the model's ratings on the corpus and the construction is described in Section 2.6.

2.2 Word Specificity Estimation: $sp(z_i, \boldsymbol{d})$

Instead of giving all words of a topic the same weight when estimating $p(z_i|\boldsymbol{d})$, we account for their ability to identify a topic by classifying words w by their posterior topic entropy $H[Z_w]$ with $Z_w \sim p(z|w)$ and their information content given the topic, defined as $-\log p(w|z)$. We transformed the entropy values to a linear scale by calculating $2^{H[Z_w]}$, which we have found to better discriminate the cases.

Words with low entropy are closely tied to only few topics and should therefore contribute more to their topic's proportions. We further discriminate them into *signal* and *specific* words through their information content. A *signal* word has low information content for a given topic and thus is likely to occur whenever the topic is present in a document. On the contrary, a *specific* word has a high information content and identifies very specific stories for a given topic, because it is rather uncommon for the topic.

We found that the main topics of an article often consist of both kinds of words while less prominent topics tend to contain many signal words. To decrease their influence, we use $sp(z_i, \boldsymbol{d})$ as defined in Equation 4 to estimate the proportion of topic z_i. In Equation 5, $\pi_{z_i}(\boldsymbol{d})$ denotes the projection of the document on the words assigned to topic z_i, while $freq(w_j)$ represents the frequency of the word w_j.

$$sp(z_i, \boldsymbol{d}) = \frac{specificity(z_i, \boldsymbol{d})}{\sum_{j=1}^{k} specificity(z_j, \boldsymbol{d})} \tag{4}$$

$$specificity(z_i, \boldsymbol{d}) = \sum_{\substack{w_j \in \\ \pi_{z_i}(\boldsymbol{d})}} freq(w_j) \cdot \frac{-\log(p(w_j|z_i))}{2^{H[Z_{w_j}]}} \tag{5}$$

2.3 Topic Dissimilarity Determination: $dis(z_i, z_j)$

Similar to [3], we found that topic similarity is best measured by topic co-occurrences across documents. Inspecting the cosine similarity for all topic pairs as depicted in Figure 1, we found that most pairs of topic vectors were almost orthogonal. When evaluating dissimilarity functions, we found that $dissim_{linear}(z_i, z_j) = 1 - sim(z_i, z_j)$ made a document's overall dissimilarity score highly dependent on its topic proportions, while $dissim_{inverse}(z_i, z_j) = 1/sim(z_i, z_j)$ was problematic with unimportant topics that have small proportions. As topics with small proportions were fairly common and highly dissimilar pairs among these might randomly occur, overall scores could become misleading.

Thus, we use *normalized pointwise mutual information* (NPMI), which measures how well two outcomes, here the two topics z_i and z_j, are determined by each other. It is defined in Equation 6, where $p(z_i, z_j) = \sum_{\boldsymbol{d} \in D} p(z_i|\boldsymbol{d}) \cdot p(z_j|\boldsymbol{d}) \cdot p(\boldsymbol{d})$. To quantify dissimilarity, we construct $dis(z_i, z_j)$ as given in Equation 7. It interpolates the NPMI values to a scale of 0 to 1 to reflect the fact that the most similar topic pair does not contribute to the unexpectedness potential of an article.

$$npmi(z_i, z_j) = \log \frac{p(z_i, z_j)}{p(z_i) \cdot p(z_j)} / (-\log p(z_i, z_j)) \tag{6}$$

$$dis(z_i, z_j) = \frac{npmi(z_i, z_j) - \min_{z_a, z_b \in Z} npmi(z_a, z_b)}{\max_{z_a, z_b \in Z} npmi(z_a, z_b) - \min_{z_a, z_b \in Z} npmi(z_a, z_b)} \tag{7}$$

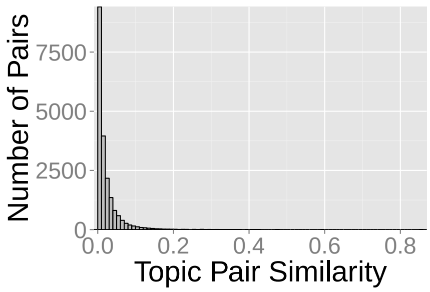


Fig. 1: Histogram of topic pair cosine similarity.

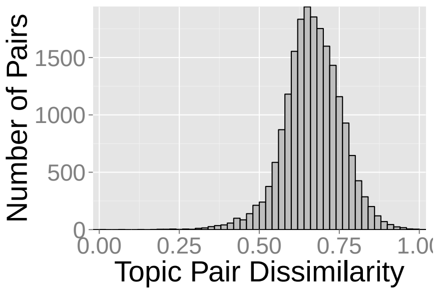


Fig. 2: Histogram of *dis* normalized to a $[0; 1]$ scale.

The resulting histogram when calculating $dis(z_i, z_j)$ on the corpus' topic pairs is depicted in Figure 2 and resembles a bell-shaped curve. Using this formulation, we experienced fewer problems with small topics while enjoying a sound information-theoretic foundation that better fits the probabilistic LDA model than a cosine-based formulation.

2.4 Limitation of Small Topics' Influence: $c(z_i, z_j, \boldsymbol{d})$

In the entire corpus, 80 percent of the the word-specific topic proportions in documents $sp(z_i, \boldsymbol{d})$ was determined by an average of 3.69 topics per document, while the remaining 20 percent consisted of 7.82 "Small topics", displaying moderate relative, but small absolute portions that influencing the model significantly.

Thus, we restrict the unexpectedness score calculation to the largest topics that make up 80 percent of a document, denoted as $Z_{Main}(\boldsymbol{d})$ and quantify our confidence whether a topic z_i in a document $\boldsymbol{d}$ can be recognized by a reader in shorter articles by $rec(z_i, \boldsymbol{d})$ as specified in Equation 8. $\pi_{z_i}(\boldsymbol{d})$ is the projection of the document on the words assigned to topic z_i. The logarithm ensures that the bias towards long document topics is less extreme. Finally, we define the confidence of a topic pair in a document $c(z_i, z_j, \boldsymbol{d})$ as the harmonic mean of their *rec* values as specified in Equation 9.

$$rec(z_i, \boldsymbol{d}) = log(||\pi_{z_i}(\boldsymbol{d})||_1 + 1) \tag{8}$$

$$c(z_i, z_j, \boldsymbol{d}) = \frac{2 * rec(z_i, \boldsymbol{d}) * rec(z_j, \boldsymbol{d})}{rec(z_i, \boldsymbol{d}) + rec(z_j, \boldsymbol{d})} \tag{9}$$

2.5 Topic Variety Normalization: $norm(\boldsymbol{d}, Z_{Main}(\boldsymbol{d}))$

According to our notion, serendipity occurs when at least one combination of unexpected topics is present. Thus, it is less important whether an article consists

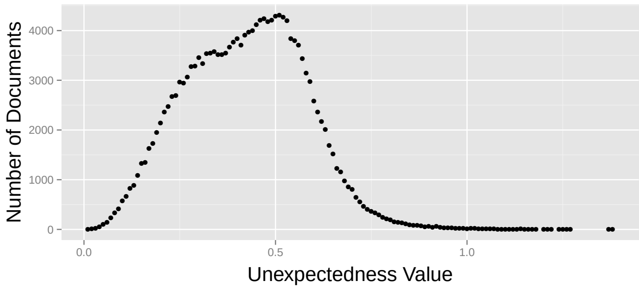


Fig. 3: Histogram of unexpectedness values for the 209,467 corpus documents. 31,968 articles with value 0 are omitted.

of more than two unexpected topics. However, by summing over all topic pairs, our model accounts for a document's topic variety, which is captured by the total number of subgroups of dissimilar topics. A larger set of main topics $Z_{Main}(\boldsymbol{d})$ implies a larger unexpectedness value when assuming similar topic proportions and equal dissimilarity values.

To account for this bias, we normalized a document by its set of main topics $Z_{Main}(\boldsymbol{d})$, i.e., $sp(z_i, \boldsymbol{d}) = \frac{1}{|Z_{Main}(\boldsymbol{d})|}$ and defined the uniform Gini Index $gini_{uniform}$ as in Equation 10 and $norm$ accordingly in Equation 11. This formulation is indifferent of the true proportions in $Z_{Main}(\boldsymbol{d})$ and keeps the model's property to account for the topic balance.

$$gini_{uniform}(\boldsymbol{d}, Z_{Main}(\boldsymbol{d})) = \sum_{\substack{z_i, z_j \in Z_{Main}(\boldsymbol{d}) \\ z_i \neq z_j}} \left(\frac{1}{|Z_{Main}(\boldsymbol{d})|} \right)^2 \tag{10}$$

$$norm(\boldsymbol{d}, Z_{Main}(\boldsymbol{d})) = \frac{1}{gini_{uniform}(\boldsymbol{d}, Z_{Main}(\boldsymbol{d}))} \tag{11}$$

2.6 Corpus Exploration

We build our LDA model using the Mallet toolkit[2] and employ standard preprocessing techniques. The distribution of unexpectedness scores from Equation 3 for our corpus are shown in Fig 3. The scores are in $[0; 1.38]$ with median 0.38 when including the zero-valued articles, which consist of a single main topic, and 0.46 without. The *long tail* of highly unexpected articles in the range $[0.8; 1.38]$ accounts for 0.78 percent of the corpus.

3 Evaluation

While our unexpectedness model can be used to rank documents, in a recommendation context, it does not consider the *source article* being read by a user who presumably shows interest in its topics. As serendipity relies on unexpectedness

[2] http://mallet.cs.umass.edu/

and interestingness, we suppose that recommendations have to demonstrate a certain similarity to it in order to ensure the reader could also be interested in them. In this chapter, we evaluate our recommendation algorithm against a set of baseline algorithms in a final user study[3].

3.1 Ranking Algorithms

To identify an adequate combination of unexpectedness and interestingness for our serendipity model, we measured the individual influence of similarity and unexpectedness in the task of making serendipitous recommendations and added two baseline algorithms to evaluate the ranking strategies with respect to their induced serendipity. The ranking algorithms used were:

1. $rank_{\text{unexp}}$ ranked according to the score u of our unexpectedness model.
2. $rank_{\text{cosine}}$ ranked articles by their cosine similarity to the source article based on tf-idf document vectors.
3. $rank_{\text{diversity}}$ ranked articles by their diversity according to the model by Bache et al. [3] that we used as the basis of our work.
4. $rank_{\text{dissimilarity}}$ ranked articles by their topical dissimilarity with the source article, quantified by the Kullback Leibler divergence of the the two articles' topic distributions as given by the LDA model.
5. $rank_{\text{serendipity}}$ re-ranked articles from $rank_{\text{unexp}}$ and $rank_{\text{cosine}}$ with the boosting strategy of Section 2.

As $rank_{\text{unexp}}$ and $rank_{\text{diversity}}$ do not consider the source article, many completely unrelated and thus probably very uninteresting articles were ranked prominently and would affect the evaluation. We therefore introduced a first step in which completely unrelated articles are excluded from the different rankings. We quantify the relatedness by calculating the cosine similarity to the source article with tf-idf document representations and select a similarity threshold of 0.2 to avoid irrelevant articles while keeping the set of retrieved articles large enough so that different re-ranking strategies could still be accurately discerned.

3.2 User Study

We evaluated the different re-ranking strategies in a user study. Six source articles were randomly selected, and for each article, the five highest ranked article recommendations from each re-ranking strategy were collected and presented to the participants in a random order. Due to articles being recommended by more than one strategy, these unions contained 14 to 19 recommendations.

Each article was presented with its headline, publication date, and categorical classifiers provided with the corpus. To make the evaluation task less time-consuming, a short abstract was provided, along with the choice to display the

[3] All data from the evaluation can be found at `https://hpi.de//en/naumann/projects/knowledge-discovery-and-mining/serendipity.html`

entire article. As the corpus-supplied abstracts were only given for approximately one third of the data set, we applied the extractive summarization algorithm *KL-Sum* [8] that employs Kullback Leibler divergence to select sentences with the most similar word distribution to that of the original document. The resulting abstracts were manually inspected and found to have an overall similar summary quality except for one thread, where the articles recommended by $rank_{\text{unexp}}$ were substantially longer than the rest, resulting in a worse quality of the extracted summaries. We therefore removed this news thread from the evaluation.

As a response format, we used an integer scale ranging from 1 to 5 and displayed the options *Strongly Disagree* and *Strongly Agree* at the two extremes. In this way, two adjacent response options were equidistant and parametric statistics like means or variance could be calculated. For each article, we displayed the following two statements:

S1: This article is relevant regarding the source article.
S2: I am positively surprised by this article. I am glad I found it.

S1 regards the relevance of a recommended article with respect to the source article, while **S2** expresses the perceived serendipity when encountering an article. As the term *serendipity* represents a complex concept, we avoided the term and described it as *positive surprise*, expressing that the article is unexpected, but nevertheless useful to the reader.

3.3 Results

Different participants tend to give ratings in different breadth and use different lower and upper bounds for bad and good recommendations. We thus employed *z-score normalization* [6], a common approach for normalizing ratings from different users to a common scale in a recommendation setting. Note that this normalization was carried out separately for each of the two statements, because they might display different rating behavior. The resulting normalized scores expressed by how many standard deviations the original scores deviated from the mean. Accordingly, a positive score indicated an above average answer.

To evaluate the serendipity of the algorithms, statement **S2** had to be assessed. The mean of the participant mean values expressed the average preference of an algorithm's serendipity by any participant. To compare the difference between two algorithms, the mean ratings of all participants were compared by a paired t-test at significance level 0.05. We were most interested in comparing $rank_{\text{unexp}}$ with the other three baseline approaches and thus Bonferroni-corrected the significance level to $\frac{0.05}{3} \approx 0.017$.

27 people took part in the study, resulting in 49 pairs of participants and source articles. All participants had a background in natural sciences and academia. The number of participants per source article ranged between 4 and 13, while

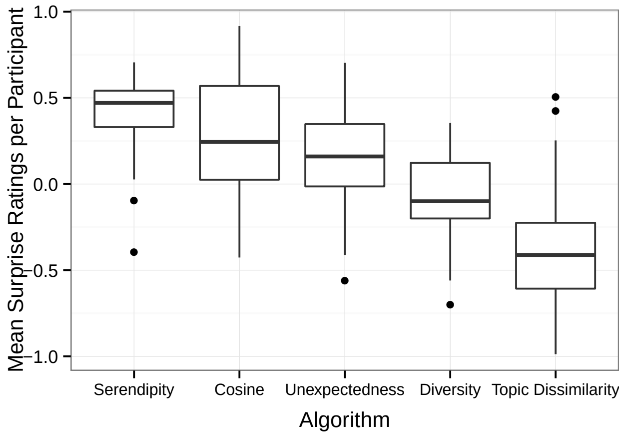


Fig. 4: Range of participant mean ratings grouped by algorithm. Each column describes the mean ratings of all 27 participants for a specific re-ranking algorithm.

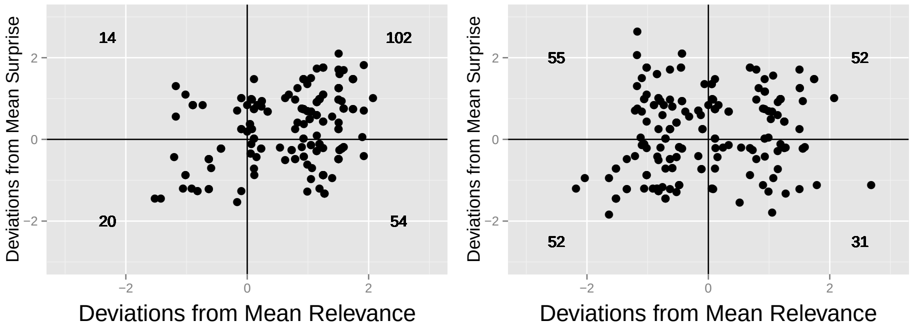
most participants rated the recommendations for a single source article. However, three participants also rated the recommendations for five or six source articles.

We computed the box-plots of algorithm means for surprise ratings (statement **S2**) across all participants for each of the methods in Figure 4 and could verify the high quality of the serendipity model. We also found that although $rank_{\text{cosine}}$ with a median value of 0.24 and mean value of 0.25 performed slightly better than $rank_{\text{unexp}}$ with mean and median value of 0.16, the difference was statistically not significant ($p > 0.1$). Furthermore, the mean and median values of $rank_{\text{unexp}}$ were significantly higher than those of $rank_{\text{diversity}}$ and $rank_{\text{dissimilarity}}$ ($p < 0.01$). We concluded that our unexpectedness function generates more surprising recommendations than these two algorithms. The comparison of $rank_{\text{cosine}}$ and $rank_{\text{unexp}}$ with $rank_{\text{serendipity}}$ is described in the next Section.

3.4 A Combined Serendipity Model

While $rank_{\text{cosine}}$ generated the most positively surprising, and therefore potentially serendipitous, articles compared to $rank_{\text{unexp}}$, the latter algorithm also generated better than average surprising recommendations. The mean Spearman's rank correlation between the recommendation rankings for all five source articles was 0.09, which means that rankings of both strategies were almost uncorrelated, indicating that serendipity occurs among articles that are highly similar as well as among articles that are less similar. This became obvious when we put the ratings for **S2**, capturing an article's serendipity, in relation to the ratings for **S1**, capturing an article's relevance to the source article, and obtained Figures 5a and 5b.

For $rank_{\text{cosine}}$, most of the positively surprising articles are concentrated in the similarity range $[0.4; 0.7]$, while they are in the range $[0.3; 0.4]$ for $rank_{\text{unexp}}$.



(a) (b)

Fig. 5: Comparison of the distribution of ratings in the two dimensions of relevance (statement **S1**) and positive surprise (statement **S2**) for $rank_{\text{cosine}}$ (a) and $rank_{\text{unexp}}$ (b). For each quadrant, the absolute number of ratings is given as some points overlap.

Using this knowledge, we combined $rank_{\text{unexp}}$ and $rank_{\text{cosine}}$ to a joined serendipity ranking $rank_{\text{serendipity}}$ by constructing a boosting algorithm that estimates the likelihood that each article recommended by $rank_{\text{unexp}}$ or $rank_{\text{cosine}}$ will likely have a positive surprise rating, based on the ranking algorithm as well as on the similarity between recommended and source article.

We evaluated the five highest ranked recommendations from $rank_{\text{serendipity}}$ by ten-fold cross validation on the set of pairs of participants and chosen source articles. For the evaluation of those recommendations, the normalized surprise ratings were determined and aggregated per participant into a mean value. Over all participants' mean ratings, $rank_{\text{serendipity}}$ achieved a median surprise rating of 0.48 and mean surprise rating of 0.41. According to a paired t-test, this was significantly different from $rank_{\text{unexp}}$'s mean ($p < 0.01$). A t-test with the mean of $rank_{\text{cosine}}$, 0.25 showed results in the range $[0.1; 0.05]$, which we did not regard as significant.

Figure 6 shows the distribution of ratings across the two dimensions of the stimuli **S1** and **S2** for recommendations made by $rank_{\text{serendipity}}$. Compared to the distributions in Figure 5, $rank_{\text{serendipity}}$ successfully recommended serendipitous articles that had high and low relevance ratings. In general, the overall number of above-average serendipity ratings increased from 116 in Figure 5a and 107 in Figure 5b to 130. We furthermore noticed that the number of articles that were neither relevant nor serendipitous stayed on the level of $rank_{\text{cosine}}$. $rank_{\text{serendipity}}$ therefore successfully combines the benefits of both approaches. As $rank_{\text{serendipity}}$ was created based on the evaluation data, part of future work is to assess whether its validity holds for a different set of articles and participants.

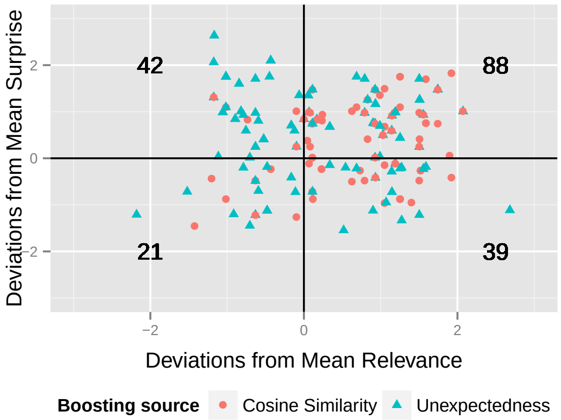


Fig. 6: Comparison of the rating distribution for the boosting algorithm $rank_{\text{serendipity}}$ in the two dimensions of relevance (according to statement **S1**) and serendipity (according to statement **S2**). For each quadrant, the absolute number of ratings is given.

This approach has to be taken with caution, because an algorithm shaped in retrospect to fit collected ratings easily overfits the data and does not generalize. Thus, the purpose of our considerations is to demonstrate that both algorithms can be synergistically combined and to outline a direction for future efforts that can be taken to develop a combined approach for the general case.

4 Related Work

Related work that aims to improve recommendations can be classified into *diversity*, *novelty*, and *serendipity*.

Diversity and Novelty. Diversity can be subdivided into two concepts: Firstly, *individual diversity* concerns the diversity of single recommendation lists. [14], target intra-list diversity using topical data and a greedy re-ranking procedure that focuses on items most dissimilar to previous items. Secondly, *aggregate diversity* concerns the diversity of all items a system recommends to its entire user base. [1] re-ranked recommendations from user-based collaborative filtering by their number of ratings to suggest less popular items.

The authors of [9] regard items that are dissimilar to the user's taste as novel and recommend diverse items, assuming that a diverse recommendation list also contains items that are novel to the user. In the field of music recommendation, the authors of [5] find that content-based systems are better at recommending novel items, as collaborative filtering systems are drawn to more popular items. However, these results do not imply that content-based systems are generally better suited to make novel recommendations.

Serendipity. Previous work on serendipity has focused on approaches that exploit a user's previous experiences with the system to induce serendipitous experiences and has been applied to various domains, such as artwork or music recommendation. For example, the authors of [4] used a lazy random walk algorithm on entities extracted from sources of user-generated content to generate serendipitous results.

A content-based recommender for artworks based on textual descriptions is built in [10] and uses a Naive Bayes classifier based on user feedback that models whether a user might like or dislike a document. Serendipitous documents are identified as those for which the classifier is most uncertain. Further, in the domain of music recommendation, the authors of [13] identify clusters of musicians that a user likes and try to recommend those musicians that belong to clusters yet unexplored by the user. A user-independent model for *general unexpectedness* of TV programmes based on word co-occurrence of their textual descriptions is presented in [2]. The serendipity model requires a clustering of items known by the user and is based on general unexpectedness as well as the distance of an item to the user-specific item clusters.

Many approaches use clustering to determine items users likely know; we also employ a soft clustering of news articles in form of topic modeling. In contrast to prior work, we propose a user-independent, content-based model that infers the unexpectedness of an article's topics from the item corpus instead of a user's preference ratings. While work on textual data of different domains exists, we address the problem for news articles.

Evaluating serendipity. Little work exists on the evaluation of serendipity in user studies. A general survey framework for recommendation systems is developed by [11]. While it does not include serendipity, the authors note that the distinction of serendipity and novelty may be confusing for participants when evaluating both. In [12], different movie recommendation algorithms are evaluated regarding aspects like novelty and serendipity by a user study. The authors note the difficulty of evaluating serendipity due to its complex definition. In our work, we employ user studies for exploration as well as evaluation, because objective evaluation metrics that are universally applicable do not yet exist.

5 Conclusion

We presented a purely content-based algorithm that recommends serendipitous news articles based on an unexpectedness model of topic combinations in articles and a traditional cosine-based similarity model. By combining both models, we were able to incorporate the advantages of both and offer users a wide variety of serendipitous articles. The unexpectedness model currently focuses only on the dissimilarity of latent topics in documents. Incorporating further content-based features, e.g., named entities, authors, publication date, or even explicit user interest captured in user profiles could substantially increase the quality of recommendations.

References

1. G. Adomavicius and Y. Kwon. Improving aggregate recommendation diversity using ranking-based techniques. *IEEE Transactions on Knowledge and Data Engineering*, 24(5):896–911, 2012.
2. T. Akiyama, K. Obara, and M. Tanizaki. Proposal and evaluation of serendipitous recommendation method using general unexpectedness. In *Workshop on the Practical Use of Recommender Systems, Algorithms and Technologies*, page 3, 2010.
3. K. Bache, D. Newman, and P. Smyth. Text-based measures of document diversity. In *Proceedings of the ACM International Conference on Knowledge Discovery and Data Mining*, KDD, pages 23–31, 2013.
4. I. Bordino, Y. Mejova, and M. Lalmas. Penguins in sweaters, or serendipitous entity search on user-generated content. In *Proceedings of ACM International Conference on Information and Knowledge Management*, CIKM, pages 109–118, 2013.
5. O. Celma and P. Herrera. A new approach to evaluating novel recommendations. In *Proceedings of the ACM Conference on Recommender Systems*, RecSys, pages 179–186, 2008.
6. C. Desrosiers and G. Karypis. A comprehensive survey of neighborhood-based recommendation methods. In *Recommender Systems Handbook*, pages 107–144. Springer, 2011.
7. M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: Evaluating recommender systems by coverage and serendipity. In *Proceedings of the Fourth ACM Conference on Recommender Systems*, RecSys, pages 257–260, 2010.
8. A. Haghighi and L. Vanderwende. Exploring content models for multi-document summarization. In *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 362–370. Association for Computational Linguistics, 2009.
9. N. Hurley and M. Zhang. Novelty and diversity in top-n recommendation – analysis and evaluation. *ACM Transactions on Internet Technology*, 10(4):14:1–14:30, 2011.
10. L. Iaquinta, M. De Gemmis, P. Lops, G. Semeraro, M. Filannino, and P. Molino. Introducing serendipity in a content-based recommender system. In *Eighth International Conference on Hybrid Intelligent Systems*, pages 168–173, 2008.
11. P. Pu, L. Chen, and R. Hu. A user-centric evaluation framework for recommender systems. In *Proceedings of the ACM Conference on Recommender Systems*, RecSys, pages 157–164, 2011.
12. A. Said, B. Fields, B. J. Jain, and S. Albayrak. User-centric evaluation of a k-furthest neighbor collaborative filtering recommender algorithm. In *Proceedings of the Conference on Computer Supported Cooperative Work*, CSCW, pages 1399–1408, 2013.
13. Y. C. Zhang, D. O. Séaghdha, D. Quercia, and T. Jambor. Auralist: Introducing serendipity into music recommendation. In *Proceedings of the ACM International Conference on Web Search and Data Mining*, WSDM, pages 13–22, 2012.
14. C.-N. Ziegler, S. M. McNee, J. A. Konstan, and G. Lausen. Improving recommendation lists through topic diversification. In *Proceedings of the International Conference on World Wide Web*, WWW, pages 22–32, 2005.