Computational Surprise, Perceptual Surprise, and Personal Background in Text Understanding

Xi Niu
University of North Carolina at Charlotte
Charlotte, North Carolina
xniu2@uncc.edu

Fakhri Abbas
University of North Carolina at Charlotte
Charlotte, North Carolina
fabbas1@uncc.edu

ABSTRACT

The concept of surprise has special significance in information retrieval in attracting user attention and arousing curiosity. In this paper, we introduced two computational measures of calculating the amount of surprise contained in a piece of text, and validated with the perceived surprise by users with different background knowledge expertise. We utilized a crowdsourcing approach and a lab-based user study to reach a large amount of users. The implication could be used to propose or refine future computational approaches to better predict human feeling of surprise triggered by reading a body of text.

CCS CONCEPTS

• **Information systems → Users and interactive retrieval**;

KEYWORDS

surprise; computational approach; personalization; crowdsourcing

ACM Reference Format:
Xi Niu and Fakhri Abbas. 2019. Computational Surprise, Perceptual Surprise, and Personal Background in Text Understanding. In *2019 Conference on Human Information Interaction and Retrieval (CHIIR '19), March 10–14, 2019, Glasgow, United Kingdom.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3295750.3298963

1 INTRODUCTION

Emotions play an important role in effective understanding of text, and therefore have attracted the interest of researchers in information retrieval. Recently there is growing research interest in sentiment analysis, which uses computational approaches to detect sentiments, emotions, feelings, or attitudes in a given body of text. While sentiment analysis is understanding the text authors' emotions, IR researchers also care about the readers' feelings that could be triggered by reading a piece of text, in order to better serve the retrieval purpose.

Among different kinds of emotions, surprise has special significance in information retrieval because natural human information discovery processes are full of surprises, from finding a bizarre movie at Netflix to discovering X-rays in scientific breakthroughs. Previous studies have shown that surprising or unexpected discovery attracts user attention and may arouse pleasant feelings, such as interest, like, and curiosity (e.g., [5, 6]). The explanation is in neuro-science, where it is believed seeking surprise is a human trait. Only the surprising signal at one stage is transmitted to the next stage [8]. Hence, human sensory cortex has adapted to predict and downplay the expected regularities of the world [4, 7], focusing instead on events that are unpredictable or surprising. However, compared to some other user feelings such as relevance, interest, and satisfaction, surprise has not received much attention in IR community, probably due to the vague and elusive nature of the feeling. Surprise by nature contradicts intention and control, which are incompatible with modern IR approaches that highly rely on computing.

In this paper, we introduce our effort in constructing computational measures of surprise and validating them with user perceptions. We call the former computational surprise and the latter perceptual surprise. We also examine the personalization factor of background knowledge expertise and its relationship with perceptual surprise with the hypothesis that it is harder to surprise a knowledgeable person. We use a corpus of health news as the experiment dataset to investigate those research problems. The contribution of this research is three-fold. Most importantly, the research compares computational surprise by machine and perceptual surprise by human. The implication could be used to refine the future computational approaches to better predict human feeling of surprise. Second, taking the individual background knowledge into consideration contributes to better understanding of the personalization factor for surprise, which further informs the construction of computational models of surprise. Third, this study also contributes a human-labeled text corpus collected from a crowdsourcing platform and a lab-based user study. As we know, so far such kind of datasets that contain human annotations and evaluations on surprise is not available.

2 COMPUTATIONAL MEASURES OF SURPRISE

We have introduced our surprise definition and various computational measures in our previous papers [removed for review]. Below is a brief summary of our previous work that is needed in this study.

We have defined that the amount of surprise as the distance to the expectation, represented as in Equation 1:

$$surprise = dist(expectation) \quad (1)$$

CHIIR '19, March 10–14, 2019, Glasgow, United Kingdom

ACM ISBN 978-1-4503-6025-8/19/03…$15.00
https://doi.org/10.1145/3295750.3298963

Through this equation, we have converted the problem of modeling surprise to a problem of modeling expectation. Based on this definition, we will briefly summarize two of our proposed measures, as documented in [removed for review].

2.1 Topic-MI and Theme-KL

In our first approach, we view each news article as "a bag of co-occurring topics", where topics are the labels assigned to an article by experts or users. The expectation of seeing an article is modeled as the expected likelihood of a particular bag of co-occurring topics in the corpus, represented as the multiplication of the individual likelihood of each topic contained in the article, i.e., p(t_1)*p(t_2)...*p(t_n). The actual or observed likelihood for an article, however, is the joint likelihood of the topic combination, i.e., p(t_1, t_2,...,t_n). The difference between the observed likelihood and the expected likelihood reflects the amount of surprise, represented as the negative of the log ratio to discount the very large difference, as in Equation 2:

$$s_1 = -log_2 \frac{p(t_1, t_2, ..., t_n)}{p(t_1)p(t_2)...p(t_n)} \quad (2)$$

where s_1 represents the surprise score of the article calculated by this method. This surprise calculation is a variation of an established metric in text mining field called Mutual Information (MI) [3]. For later reference, we label method as Topic-MI.

Going deeper or more fine-grained than topics, in second approach, we view each article as "a bag of co-occurring latent themes". Probabilistic topic modeling [1] is a set of algorithms that discover the latent themes that run through words and how these themes are connected. LDA (Latent Dirichlet Allocation) [2] is a popular example of these algorithms and is applied in this study. According to LDA, each article is generated by choosing the latent themes z_i probabilistically and then for each latent theme choosing the word w_i probabilistically, represented as:

$$p(d) = \Pi_{i=1}^{k} p_\alpha(z_i) \Pi_{j=1}^{\theta} p_\beta(w_j|z_i) \quad (3)$$

where p_α is the distribution of the latent themes in an article, z_i is the latent theme i, $p_\beta$is the distribution of the words for the latent theme z_i, and w_j is the word j. In fact, the generative model is the likelihood of observing such an article with k latent themes. We apply this model such that the expected likelihood of seeing a typical article will be the likelihood of observing an "average" article by averaging the distributions of all articles in the main topic that this article addresses. Each individual article's divergence from this expectation is that article's degree of surprise. We use the Kullback Leibler (KL) divergence as the divergence measure. The surprise score s_2 is calculated as Equation 4, where p_α is the distribution of the latent themes in an article, q is the distribution of a typical article, and i is the index of the latent themes. We label this approach as Theme-KL.

$$s_2 = KL(p_\alpha, q) = \sum_{i=1}^{k} p_{\alpha_i} log_2 \frac{p_{\alpha_i}}{q_i} \quad (4)$$

2.2 Topic-MI and Theme-KL on the Health News Corpus

To apply Topic-MI and Theme-KL, we have scraped the health news articles from Medical News Today (MNT) since its launch in 2003 to the present. The corpus contains 268,850 articles, classified into 135 health topics, such as diabetes, heart disease, anxiety, women's health, by health professionals working with MNT. Most articles have multiple topic labels.

For the Topic-MI approach, these 135 topic labels were leveraged as the topics (t_i) as in Equation 2. The value of s_1 for each article was calculated using Python's *math* and *sklearn* packages. For the Theme-KL approach, a high efficiency topic modeling tool, *gensim*, a Python library, was used for the LDA analysis and calculating s_2 for each article.

To demonstrate the computational result of the surprise scores, let us take the topic *cancer* as an example. Figure 1 presents the cumulative distribution of the Z-scores of s_1 and s_2 respectively of the 27,760 articles that invovle the topic *cancer*. For the s_1 distribution in Figure 1(a), a Z-score at the right side (above the average 0, highlighted in gray) indicates a topic combination with cancer potentially surprising, co-occurring less often than expected by chance. One such example is the combination of *cancer*, *veterinary*, *drug resistance*, and *public health*, as shown as one example in Table 1. For the s_2 distribution in Figure 1(b), a Z-score at the right side indicates a larger divergence from an average article. This suggests a potentially surprising latent theme distribution, such as an article talking about a new brochure that helps patients understand clinical trials, as shown in the example table (Table 1). Both curves are roughly parallel to a normal distribution, suggesting that most articles are centered around the average level of surprise. The distribution for the Topic-MI approach is skewed to the higher end whereas that of the Theme-KL approach to the lower end, implying the Topic-MI approach is more generous in giving high scores whereas the Theme-KL approach is more selective in identifying highly surprising articles. Table 1 lists some examples of the most and the least surprising articles based on s_1 and s_2. We will validate these scoring mechanisms with user perceived surprise.

3 HUMAN PERCEPTION COLLECTION

No matter which approach (Topic-MI or Theme-KL) we use, the surprise is a "database" surprise obtained through mining a large corpus, which represents the society's collective knowledge. However, for individuals, their feelings of surprise are different because of their background difference. We will use an interactive process to gather information about user background and their perceptual amount of surprise to investigate how the same amount of "database" surprise is felt by different people with different backgrounds. Our assumption is that it is harder to surprise a knowledgeable person than a novice. For example, a news article about male breast cancer would be surprising to most of people but not necessarily so for those who know that it is not only women who can develop breast cancer and males can too.

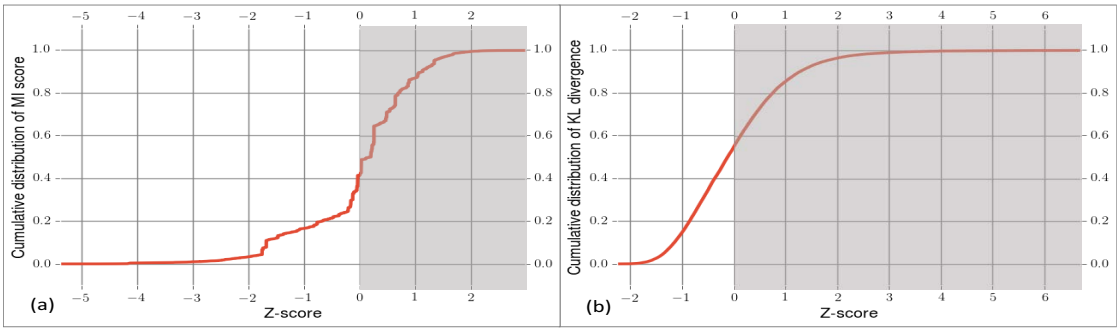

Figure 1: Topic-MI (a) and Theme-KL (b) distributions for all the *cancer* articles

Table 1: Article examples based on s_1 and s_2

Example Article Titles for the MI Approach	Topic Labels	s_1	Z-score
Are routine pelvic exams 'more harm than good' for healthy women?	Cancer, Women's Health, Infectious Disease	5.13	2.47
Hospital for humans to offer radiotherapy for animals criticized	Cancer, Veterinary, Drug Resistance, Public Health	5.10	2.46
Father's Day gift: encourage dad to go for prostate screening	Cancer, Prostate Cancer, Urology, Men's Health	-12.02	-4.13
University of Southern California study explains major cause of drug-resistance in chronic myeloid leukemia	Cancer, Drug Resistance, Lymphoma	-12.44	-4.29
Example Article Titles for the KL Approach	**Topic Labels**	**s_2**	**Z-score**
Fostering cancer research - Miss America, Dan Haney and Randy 'Duke' Cunningham win awards	Cancer	2.18	6.70
New brochure helps cancer patients understand clinical trials	Cancer	2.10	6.33
New 'targeted' treatments improve colon cancer survival rates	Cancer, Colorectal Cancer, Gastrointestinal	0.21	-2.10
Enzyme responsible for brain tumors discovered	Cancer, Biology, Neurology	0.21	-2.11

* the gray area indicates surprising articles whereas the white area indicates non-surprising ones

3.1 The Crowdsourcing Approach

To reach a large amount of users, we used Amazon Mechanical Turk[1] (MTurk), the well-known crowdsourcing platform. We posted 500 human intelligence tasks (HITs). Each HIT presented a health news article to the recruited workers. The workers needed to offer their ratings on 5-point Likert scales on three questions: 1) how surprising they think the article is, 2) how much they like the article, and 3) how familiar they are with the topics contained in the article. The surprise ratings are the core information we want to collect from the workers. The reason to have this "like" question is that we need to understand whether the encountered surprise is favored by the person. Favorable surprise instead of random or negative surprise is the further target of our future computational approaches or machine learning models. There are many implementation ways to collect the information about a person's background knowledge. In this study, we have adopted a simple way of asking their familrity on the article topics, for the proof-of-concept purpose.

For the 500 HITs, we hand-selected 500 articles with varying scores of s_1 and s_2 from five popular health topics: *Diabetes*, *Depression*, *Nutrition*, *Cancer*, and *Children's Health*. For each topic, we hand-selected 100 articles with the focus on everyday reading, avoiding those research-oriented articles with chemistry symbols or medical terminologies. Although we selected five topics, most of the 500 articles also contain other topics outside the five topics. For each HIT, we have recruited three different workers. Therefore, we have collected 1,500 (500 x 3) cases of user ratings.

3.2 Lab-Based User Study

To supplement the user ratings collected from MTurk, we have made use of the user ratings collected from one of our previous user studies [removed for review]. The procedure is summarized briefly in below. Thirty graduate students were recruited into our lab for a one-hour study. None of them had completed any formal medical education. After the introduction and obtainment of consent, each participant was recommended 50 - 100 articles with varying scores of s_1 and s_2. These articles were separated into 10 sessions. They were encouraged to click on whatever articles they would like to read. For each clicked article, they were required to provide their ratings on 5-point Likert-scales on same three questions as in the crowdsourcing study: whether the article content is surprising, whether they like the article, and how familiar they are with the topics contained in the article. As the result, we have collected 497 cases of user ratings.

[1]https://www.mturk.com/

4 RESULTS

For the 1,997 cases combined, the distribution of the surprise, like, and familiarity ratings are presented in Figure 2. The surprise ratings are rather evenly distributed across the five categories. As to the like ratings, users are very generous in giving their ratings, with 4 as the most frequent rating. In terms of the self-reported familiarity level on the main topic of the article, most of them rated themselves as having a medium familiarity level. The familiarity ratings follow a normal distribution.

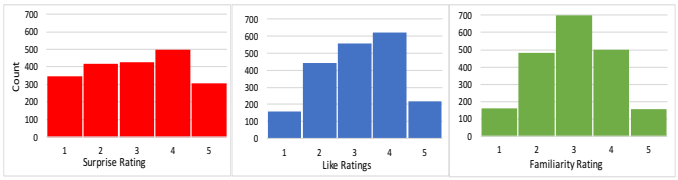

Figure 2: The distribution of the surprise ratings, like ratings, and familiarity ratings

4.1 Computational Surprise vs. Perceptual Surprise

Pairwise Pearson correlations between s_1, s_2, and the surprise ratings are summarized in Table 2. The non-significant correlation between s_1 and s_2 (r = 0.13, p = 0.2054) suggests that the two approaches do not agree with each other in calculating the amount of surprise an article contains. However, the correlations between s_1 and the surprise ratings as well as s_2 and the surprise ratings are both significant (r = 0.46, p < 0.0001; r = 0.33, p = 0.0107), meaning both computational approaches are able to capture what the users think surprising to some degree, but in different aspects. MI-Topic is better aliened with what users think, given s_1's higher correlation coefficient.

When we manually checked into the contents of the surprising articles identified by each approach, we found that the Topic-MI approach is to find the articles with rare topic co-occurrence whereas the the Theme-KL approach was good at finding "atypical" articles, such as ones on government policy, insurance, academic conference announcements, etc. Such atypical content was captured as peripheral latent themes.

We are also interested in the correlation between the computational surprise and the perceived favorable surprise, implemented as the sum of the surprise rating and the like rating. The reason is that favorable surprise, not negative surprise or random surprise, is the ultimate goal for the future retrieval models. As shown in Table 3, both correlation coefficients drop compared to Table 2, with one remains statistically significant and the other not. This means that although our computational approaches are able to find surprising contents, but not necessarily favorable surprising contents by users. Future work needs to construct separate computational models of "favor", as the content-based or collaborative filtering algorithms used in recommender systems, before applying the computational approaches of surprise, in order to restrict the search for surprise in the space of what users favor.

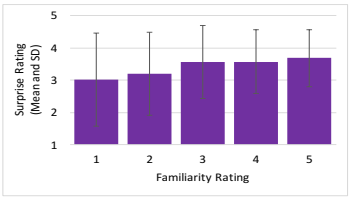

Figure 3: Surprise ratings vs. familiarity ratings

Table 2: Pearson correlation between computational surprise and perceptual surprise

	s_1	s_2	surprise rating
s_1	1		
s_2	0.13	1	
surprise rating	0.46*	0.33*	1

Notes: * denotes statistical significance at 0.05 level.

Table 3: Pearson correlation between computational surprise and perceived favorable surprise

	s_1	s_2
surprise rating + like rating	0.28*	0.17

Notes: * denotes statistical significance at 0.05 level.

4.2 Perceptual Surprise vs. Topic Familiarity

Since each news article has a main topic, on which the user has indicated their familiarity level. The articles' surprise ratings grouped by user familiarity levels are presented in Figure 3. Unexpectedly, the participants with the highest familiarity level have provided the highest average surprise rating whereas those who barely know about a topic have offered the lowest average surprise ratings. ANOVA test shows there is significant difference among these different familiarity groups (F(4, 1992) = 2.84, p = 0.0230). The standard deviations, as shown in the error bars in Figure 3, are generally decreasing from the familiarity group 1 through the group 5. The decreasing variance suggests that people with the familiarity level of 5 are more consistent in offering their ratings. The finding is different from our expectation that it would be more difficult to surprise the knowledgeable. On the contrary, the users who know more are in fact more generous in giving high ratings of surprise. It could be that those knowledgeable users are more certain about what they know and they do not know, and therefore more capable of recognizing surprise when seeing them.

5 CONCLUSION AND FUTURE WORK

We have evaluated whether the computational surprise aligns with the perceived surprise by users with different background. As the result, we find that although the two computational surprise measures do not correlate, they each moderately correlate with the user perceived surprise, capturing different aspects of what users think surprising. We unexpectedly find that the knowledgeable users perceive more surprise than the novice users, probably because they are more capable of recognizing surprise. Future work includes in constructing machine learning models, leveraging s_1, s_2, and topic familiarity as features, to predict what text will trigger the readers' feeling of surprise.

REFERENCES

[1] David M Blei. 2012. Probabilistic topic models. *Commun. ACM* 55, 4 (2012), 77–84.

[2] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[3] Gerlof Bouma. 2009. Normalized (pointwise) mutual information in collocation extraction. *Proceedings of the Conference of the German Society for Computational linguistic and Language Technology,(GSCL)* (2009), 31–40.

[4] Valentin Dragoi, Jitendra Sharma, Earl K Miller, and Mriganka Sur. 2002. Dynamics of neuronal sensitivity in visual cortex and local feature discrimination. *Nature neuroscience* 5, 9 (2002), 883–891.

[5] Kazjon Grace, Mary Lou Maher, David Wilson, and Nadia Najjar. 2017. Personalised specific curiosity for computational design systems. In *Design Computing and Cognition'16*. Springer, 593–610.

[6] Luís Macedo and Amílcar Cardoso. 2005. The role of surprise, curiosity and hunger on exploration of unknown environments populated with entities. In *Proceedings of the Portuguese Conference on Artificial Intelligence*. 47–53.

[7] Bruno A Olshausen and David J Field. 1996. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature* 381, 6583 (1996), 607.

[8] Rajesh PN Rao and Dana H Ballard. 1999. Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience* 2, 1 (1999), 79–87.