Exploding TV Sets & Disappointing Laptops: Suggesting Interesting Content in News Archives based on Surprise Estimation

Adam Jatowt[1], I-Chen Hung[2], Michael Färber[3], Ricardo Campos[4], and Masatoshi Yoshikawa[2]

[1] University of Innsbruck, Austria
adam.jatowt@uibk.ac.at
[2] Kyoto University, Japan
{ichen,yoshikawa}@i.kyoto-u.ac.jp
[3] Karlsruhe Institute of Technology (KIT), Germany
michael.faerber@kit.edu
[4] LIAAD – INESCTEC & Polytechnic Institute of Tomar, Ci2 - Smart Cities Research Center, Portugal
ricardo.campos@ipt.pt

Abstract. Many archival collections have been recently digitized and made available to a wide public. The contained documents however tend to have limited attractiveness for ordinary users, since content may appear obsolete and uninteresting. Archival document collections can become more attractive for users if suitable content can be recommended to them. The purpose of this research is to propose a new research direction of *Archival Content Suggestion* to discover interesting content from long-term document archives that preserve information on society history and heritage. To realize this objective, we propose two unsupervised approaches for automatically discovering interesting sentences from news article archives. Our methods detect interesting content by comparing the information written in the past with one created in the present to make use of a surprise effect. Experiments on New York Times corpus show that our approaches effectively retrieve interesting content.

Keywords: archival document search · interestingness · news articles

1 Introduction

Document archives, such as news articles published over past decades, are accumulations of historical records and are important for the humanities and social studies, among others [27]. Accordingly, in recent years, massive digitization efforts of archival documents have been carried out by libraries, national archives, and numerous other memory institutions. The available data is already considerably large and is continuously growing. For instance, the Chronicling America[5]

[5] https://chroniclingamerica.loc.gov/

project has over 5.2 million individual newspaper pages available for viewing and downloading that were published in the USA in the last three centuries. Likewise, Google Books project[6] scanned over 6% of books that were ever published by humanity, many of which are from quite a distant past. In the Web domain, web archives like the Internet Archive[7] are also often used by the general public. Multiple national initiatives [12] have also emerged over the years to crawl national contents. This continuous development of digital document archives allows to learn about historical events and situations directly from primary sources. Yet accessing document archives is different from using a regular search engine, and may lead ordinary users to quickly lose interest or become disappointed. It may be because of the view of history held by some as boring and irrelevant [3,25,33]. This situation calls for research in novel access approaches and retrieval methods that would be adapted to the particular characteristics of archival document collections and could engage user's attention. Such systems should increase archival collections' utility by making them more attractive and interesting to modern users. In this research, we assume in particular that interesting information from the past should contain an element of surprise. Retrieving such content from document archives could surprise and amuse readers as well as evoke their interest, as the contained information would be against the presumed expectations. Note that such information is not easy to be found using a traditional search engine as it requires considerable effort and search skills. Also, although there are websites[8] listing surprising history facts or trivia, they are always manually created.

Although a few studies on identifying content about the unexpected relationships exist, they focus on non-archival data such as Wikipedia [5,36] or current news [20]. Contents in archives have however, particular characteristics due to their age as well as different and often unknown context. In this paper, we focus on extracting sentences from news article archives based on the attributes of content interestingness such as unexpectedness/surprise and importance. We then introduce two unsupervised approaches for discovering interesting content based on these aspects. In particular, the two-layer Mutually Reinforced Random Walk (MRRW) [7] is adapted to capture the novelty and importance in a temporal document collection. The key idea is to rank highly content from the past which was important at that time, yet which is novel or surprising currently. Content importance is modeled by measuring its popularity in the past according to the assumption that popular concepts in the past have more educational value than obscure ones. The second approach involves a topic co-occurrence model used to find surprising and unexpected topic combinations that co-occurred in the past. Our experiments are performed on the New York Times news corpus [26], which contains documents from 1987 to 2007.

In general, interestingness is a complex concept with little consensus about its definition and scope. It is definitely a challenge to retrieve and recommend attractive content with an objective methodology. Still, this kind of content sug-

[6] https://books.google.com/

[7] https://archive.org/

[8] For example: https://allthatsinteresting.com/interesting-history-facts https://www.thefactsite.com/100-history-facts/ https://parade.com/1099930/marynliles/history-facts/

gestion should help increase the perceived attractiveness of heritage collections and raise their utility for average users. Successful methods developed for this purpose could be either incorporated as integral components of retrieval mechanisms in archival search engines or could be harnessed to encourage users to start using archives[9].

2 Related Works

Representing interestingness by unexpectedness. One of the main problems in finding interesting patterns or data is defining *interestingness* properly. A longtime subject of psychology and cognitive science, the feeling of interestingness was even considered an emotion in the past. Silvia *et al.* [30] and Berlyne *et al.* [4] analyzed interestingness from the viewpoint of cognitive appraisal, which is a personal interpretation of a situation and possible reactions. Within computer science related studies, interestingness was studied in the task of pattern finding in knowledge discovery systems and general databases [13,19,21,31], recommender systems [1] and computational creativity [38]. The Bayesian theory of surprise assumes measuring the difference between posterior and prior beliefs of the observer [2,15]. Based on it, Itti and Baldi [14] developed model that computes expected low-level surprise in video streams which significantly correlates with eye movements of humans watching complex videos.

Geng *et al.* [11] treated interestingness as a broad concept that possibly contains features like reliability, diversity, surprise, and more. Silberschatz *et al.* [28] focused on subjective measures of interestingness, suggesting interesting information should be unexpected and actionable. Unexpectedness was also considered crucial by Padmanabhan *et al.* [23] and Adamopoulos *et al.* [1]. Moreover, the latter introduced serendipity as one of the evaluation measures. Yannakakis *et al.* [40] believed that surprise-focused search maximizes unexpectedness and accordingly proposed a surprise-oriented search algorithm. Tsurel [37] *et al.* assumed that trivia and surprise facts arouse user interest. In line with some of these previous approaches we also model interestingness with the help of the surprise and unexpectedness aspects of information, albeit in our specific case, they arise due to time passage.

Unexpected relationship detection. Several studies focused on finding unexpected relationships between data, for example, relationships between entities, which are unexpected. Boldi *et al.* [5] and Tsukuda *et al.* [36] used the Wikipedia[10] as their underlying knowledge-base to uncover unexpected relations. Tsukuda *et al.* [36] evaluated the unexpectedness of related terms extracted from Wikipedia pages on the basis of relationships of their coordinate terms. Boldi *et al.* [5] focused on finding unexpected links within hyperlinked Wikipedia articles.

Novelty detection. Interestingess is to some degree related to novelty which should be mentioned here, too. For example, TREC challenge[11], which consists of a set of tracks and tasks, such as TREC Temporal Summarization (TempSum),

[9] One could imagine a service that automatically detects interesting sentences or headlines for broad topics and publishes them daily on web portals of underlying document archives.

[10] `https://www.wikipedia.org/`

[11] `http://trec.nist.gov/`

TREC Knowledge Base Acceleration (KBA), and TREC Novelty Track, has brought about the improvement in the novelty detection for years. Features like *sentence lengths*, *named entities*, and *opinion patterns* were used in Li *et al.* [20] to analyze and improve the novelty detection on the 2002-2004 TREC novelty tracks. Farber *et al.* [10] proposed a new semantic approach to resolve the ambiguities in the languages and extract novel and relevant information from unstructured text documents. For more information, interested readers may refer to the survey on novelty, diversity and serendipity aspects in IR [16] and in recommender systems' evaluation [29].

In general, many of the prior studies developed their methods based on hyperlinked datasets like Wikipedia, which include explicit relationships. Only few tried discovering interesting information from unstructured text. Our research focus is on documents published at different times and subject to change which is inherent in long-term document archives. To the best of our knowledge, the concept of interestingness in archival contents remains largely unexplored.

3 Proposed Approaches

In this section, we describe two novel approaches: *Topic-based Mutually Reinforced Random Walk* and *Topic Pair-based Mutually Reinforced Random Walk*. Before doing that, we first discuss the input data.

3.1 Input data

In our setting, we assume a sentence to be a retrieval unit. We focus on sentences rather than entire documents for a few reasons. First, we believe that a short but attractive content would have more chance to be read by users than longer text. One of the envisioned applications assumes embedding the automatically extracted content in online archival portals. Doing this based on the entire document may be cumbersome and less flexible. Still, the users could visit the underlying documents from where the interesting sentences were extracted by following added links, especially when headlines are used as is often done in timeline summarization research [24,34], or when snippets are used by regular search engines. Nevertheless, extending the proposed approaches to returning the entire documents should be relatively easy.

We will make use of two document collections constructed for each input query, D_{past} which represents the set of sentences from a certain time period in the past T_{past} and D_{now} which represents the sentences from the "present" denoted as T_{now} and understood as some recent time span such as the last 6 months or 1 year. Sentences from D_{now} are to be solely used as a reference to support result generation from D_{past}. Our objective is to rank sentences from D_{past} and produce interesting output with the aid of the present collection D_{now}.

3.2 Topic-based Mutually Reinforced Random Walk

We introduce here our first approach. We generate a two-layered graph G using content from D_{past} and from D_{now} for constructing the layers of the graph. Each node in the graph represents a topic inferred from the respective document collection, while the edge weights represent either similarity or dissimilarity of

topics (to be described later). In particular, we run *Latent Dirichlet Allocation* (LDA) to build topic models from the sentences of D_{past} and sentences of D_{now}.

Let us denote the layer in T_{past} as $L_{PP} = \{z_1, z_2, ..., z_i\}$, and the layer in T_{now} as $L_{NN} = \{y_1, y_2, ..., y_j\}$, where z_i and y_j indicate topics from LDA models. Note that the topics in both layers are trained separately on the corresponding datasets, so that the similarities within the two layers will be computed on different topic spaces. We do not mix the datasets when performing the topic modeling in order to determine topics specific to either time period without affecting them by the data from the other time period. $term_{zi}$ and $term_{yj}$ represent the top-scored terms in topic z_i and topic y_j, respectively, according to the determined topic models. We then compute the overlap of the top l terms of topics in order to calculate edge weights. The edge weights within each layer (P(ast) and N(ow)) are computed as follows:

$$Sim^P(z_i, z_j) = \frac{term_{zi} \cap term_{zj}}{l} \tag{1}$$

$$Sim^N(y_k, y_l) = \frac{term_{yk} \cap term_{yl}}{l} \tag{2}$$

while the edge weights between the two layers are calculated as follows:

$$DisSim(z_a, y_b) = 1 - Sim(z_a, y_b) \tag{3}$$

where $Sim(z_a, y_b)$ is calculated similarly to Eqs. 1 and 2, i.e., by measuring term overlap.

We construct such a two-layered graph to find topics that were dominating in the past, yet that are not popular in the present, hence the use of similarity for edge weights within each layer and dissimilarity for edge weights between the layers. Based on this intuition the two-layer *Mutually Reinforced Random Walk* (MRRW) [7] is executed on the graph to assign scores to each topic. MRRW is an algorithm for computing the converged scores of nodes in layered graphs. Given within-layer and between-layer edge weights, the score for each node refers to its importance within the graph computed based on external mutual reinforcement between different layers through the between-layer edges.

The scores of node sets in both layers are reinforced by the following equation:

$$\begin{cases} S_P^{(t+1)} = (1-\alpha) S_P^{(0)} + \alpha \cdot E_{PP} E_{PN} S_N^{(t)} \\ S_N^{(t+1)} = (1-\alpha) S_N^{(0)} + \alpha \cdot E_{NN} E_{NP} S_P^{(t)}. \end{cases} \tag{4}$$

Here $S_P^{(t)}$ and $S_N^{(t)}$ denote the scores of the node set in the past and present layers, respectively, at the t-th iteration. E_{NN}, E_{PP}, E_{NP} and E_{PN} are matrices with the inter- and intra-layers' edge weights. After we apply Eq. 4 to the graph, the score of a node in layer L_{PP} will become higher if the node is more similar to other nodes in this layer and more dissimilar to the nodes in the layer L_{NN}. In this equation, α, which controls the interpolation weight for the propagation part, is set to 0.9 following [7]. The algorithm runs until convergence or until the change of scores becomes very small.

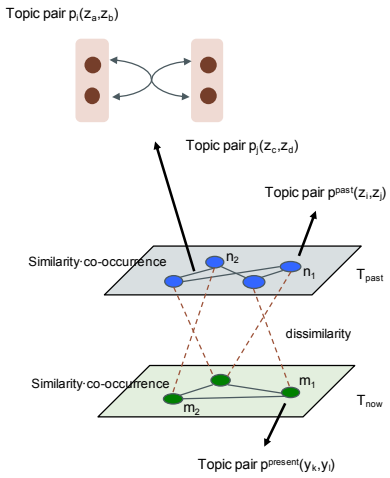

Fig. 1: The overview of the Topic Pair-based MRRW.

Afterwards, we rank the topics in L_{PP} by their computed scores. As mentioned above, the score of a past topic should be high when this topic is similar to other topics in the past while dissimilar to the topics in the present layer. For each top-ranked topic, we then retrieve the top-n sentences after computing their probability of belonging to that topic.

3.3 Topic Pair-based Mutually Reinforced Random Walk

Studies in psychology and cognitive science suggest that feeling of unexpectedness and surprise are emotional reactions when people encounter information not conforming to their stereotypical expectations [22]. We hypothesize that a sentence with a rare and uncommon combination of topics would likely be deemed unexpected or surprising. Derezinski *et al.* [9] also view topic diversity as an important element for discovering surprising documents. In this work, instead of measuring the diversity of topic distributions, we propose an approach considering uncommon topic co-occurrences to discover surprising sentences. The underlying intuition is that even if topics are not surprising, their combination could be.

For computation, we again use the two-layered graph, but now the nodes represent topic pairs (a combination of two different topics) based on the set of topics derived from each dataset. Let us denote the layer in T_{past} as $L_{PP} = \{n_1, n_2, ..., n_i\}$, and layer in T_{now} as $L_{NN} = \{m_1, m_2, ..., m_j\}$, where n is a past topic pair $p(z_i, z_j)$ and m denotes a present topic pair $p(y_k, y_l)$ as derived from LDA models. Again, topic models for either time period are trained on its corresponding data, so pair-to-pair similarities within either layer are computed over the topic set corresponding to that layer. We connect any two nodes belonging to the same layer and assign edge weights depending on the similarity and co-occurrence for each topic pair (to be described later). On the other hand,

a node pair consisting of nodes from different layers is connected by an edge whose weight represents the nodes' dissimilarity. The concept of *Topic Pair-based MRRW* is visualized in Fig. 1.

When computing the similarity between two nodes (i.e., two topic pairs), we calculate the pair-wise similarity for each possible combination of topics in the two pairs, and use the maximum similarity value as the final edge value. Same as in the above-described Topic-based MRRW, we compute the overlap of the top l topic terms to calculate the similarity and dissimilarity of two topics (Eqs. 2 and 3). We then compute the similarity between two nodes, i.e., two topic pairs in the past $p(z_a, z_b)$ and $p(z_c, z_d)$ as follows:

$$\begin{aligned} Sim^P(p(z_a, z_b), p(z_c, z_d)) = \max\{&Sim^P(z_a, z_c) \cdot Sim^P(z_b, z_d), \\ &Sim^P(z_a, z_d) \cdot Sim^P(z_b, z_c)\} \end{aligned} \tag{5}$$

while the similarity between any two nodes in the present, $p(y_a, y_b)$ and $p(y_c, y_d)$, is calculated by:

$$\begin{aligned} Sim^N(p(y_a, y_b), p(y_c, y_d)) = \max\{&Sim^N(y_a, y_c) \cdot Sim^N(y_b, y_d), \\ &Sim^N(y_a, y_d) \cdot Sim^N(y_b, y_c)\} \end{aligned} \tag{6}$$

Based on the above equations, the edge weights e within each layer are as follows:

$$e^P(n_i, n_j) = Avg\_cooc^P(n_i) \cdot Avg\_cooc^P(n_j) \cdot Sim^P(n_i, n_j) \tag{7}$$

$$e^N(m_i, m_j) = Avg\_cooc^N(m_i) \cdot Avg\_cooc^N(m_j) \cdot Sim^N(m_i, m_j) \tag{8}$$

$Avg\_cooc^P(n_i)$ and $Avg\_cooc^N(m_i)$ are the average co-occurrences of the topics in a given pair in the past and present periods, respectively. They are used here as weights which quantify the importance of topic pairs. The calculation of co-occurrence is done as follows. Sentences in both D_{past} and D_{now} are mapped to a probability distribution over topics to create a sentence-topic matrix, where each row gives a topic distribution for a sentence. The average co-occurrence of the learned topics in each time period is then computed as:

$$Avg\_cooc^P(z_i, z_j) = \frac{1}{|D_{past}|} \sum_{d_k \in D_{past}} P(z_i \mid d_k) P(z_j \mid d_k) \tag{9}$$

$$Avg\_cooc^N(y_i, y_j) = \frac{1}{|D_{now}|} \sum_{d_k \in D_{now}} P(y_i \mid d_k) P(y_j \mid d_k) \tag{10}$$

where $P(z_i|d_k)$ or $P(y_j|d_k)$ denote the probability of z_i or y_j in d_k, respectively. Finally, edge weights between the different layers are computed in a similar way to Eqs. 5 and 6 as:

$$DisSim(n_a, m_b) = 1 - Sim(n_a, m_b) \tag{11}$$

The final scores are computed by the same equation (Eq. 4) as for MRRW algorithm. After computing final scores of nodes (topic pairs), we rank the topic pairs in T_{past} by their scores, which should be higher if the topic pair is similar to the other topic pairs in the past layer while being dissimilar to the topic pairs in the present layer. For each top ranked topic pair, we then extract top-n sentences after sorting them by their probability of belonging to the corresponding topics.

4 Experimental Settings

4.1 Temporal Document Collection

We use the New York Times (NYT) News collection, which has been frequently utilized in researches of Temporal Information Retrieval [6,17] and alike. The corpus includes news articles published from 1987 to 2007. The documents contain metadata labels such as date, title, category, leading paragraph, full-text, and more. In the experiments, we divide this news archive into two parts: one from Jan. 1987 to Dec. 1989, representing past documents, and the other one containing documents published from Jan. 2005 to Dec. 2007 to represent information of the present. Naturally, the latter part is not exactly representing the "present", and is rather a compromise resulting from the lack of free datasets that would be long enough (e.g., a span of at least three different decades or more) and that, at the same time, would contain also most recent documents. When it comes to the length of time periods our choice results from the need for striking a balance between having the size of data in both the parts of the collection sufficiently large for generating topics and between maintaining a sufficiently long time gap that separates these two dataset parts. We will then process content that is roughly 30 years old as seen from now and that was published during 3 years' long time frame.

In the experiments, we consider five broad categories of concepts inspired by news categories of NYT: *Economy*, *Places*, *Politics*, *Sports*, and *Technology* as broad concepts tend to be often used by ordinary users accessing document archives [8,18,35]. Each category includes 4 general concepts resulting, in total, in 20 different concepts. Tab. 1 gives the list of categories and their concepts.

Table 1: List of categories and their concepts.

Category	Concept
Economy	Currency, Economy, Trade, Market
Places	Japan, Florida, Los Angeles, New York
Politics	Election, President, Nomination, Poll
Sports	Basketball, Team, Olympics, Sport
Technology	Machine, Computer, Plane, Technology

4.2 Preprocessing

We first find all sentences that mention the concepts using the Solr[12] search engine. We use only sentences being either the leading paragraph or the title of a document as these are most interpretable and self-contained. To ensure better understandability, we remove sentences shorter than 10 words as well as overly long sentences (longer than 50 words).

Next, we trim sentence contents by removing stopwords and punctuations using NLTK library[13]. Lemmatization is performed to handle inflections and to obtain correct base forms of words. We then use TF-IDF vectors for sentence

[12] https://lucene.apache.org/solr/

[13] https://www.nltk.org/

representation[14]. The number of topics in LDA models has been empirically set to 100 for all the approaches and the number l of top terms was also set to 100.

4.3 Baselines

Besides the two proposed approaches, we also test the following ones:

Random: We return randomly ordered sentences from the pool of candidate sentences from the past documents.

Centroid: This method ranks sentences in D_{past} by their dissimilarity to the centroid vector, which is the average TF-IDF vector of all sentences in D_{now}. It is expected to extract sentences which are less known to current users.

MRRW: This method ranks sentences by simply applying MRRW [7] on the two layers (past and present) composed of sentences treated as nodes.

Topic co-occurrence: Similarly to the proposed Topic Pair-based MRRW method, we use the concept of surprising topic pairs. However, the calculation is done without building a two-layered graph and running the random walk. To find the co-occurring topics, we use Latent Dirichlet Allocation to build a topic model over the combined sentences from D_{past} and D_{now}. Sentences in both D_{past} and D_{now} are then mapped to a probability distribution over topics $t_i \in T$. As a result, we obtain a sentence-topic matrix, where each row gives a topic distribution for a sentence. We then calculate the average co-occurrence of the learned topics in each time period using similar way as in Eqs. 9 and 10.

Topic pairs that frequently co-occur in D_{past} yet rarely in D_{now} will be ranked high by the following equation:

$$S(t_i, t_j) = \frac{Avg\_cooc^P(t_i, t_j) - Avg\_cooc^N(t_i, t_j)}{Avg\_cooc^N(t_i, t_j) + Avg\_cooc^P(t_i, t_j)} \tag{12}$$

The score of a sentence is computed by aggregating the scores of the probability of different topic pairs in the sentence. The top n sentences are then retrieved for each top-ranked topic pair same as in Topic Pair-based MRRW method.

4.4 Data Annotation

We use Figure Eight[15], a popular crowdsourcing platform to evaluate the results. We first pooled the top 15 results for the 20 queried concepts for each of the 6 tested methods[16]. This resulted in an evaluation dataset consisting of 1,800 sentences from the New York Times collection that were published between 1987 and 1989. Judges were then asked to assess the sentences based on their interestingness and surprise, and give scores ranging from 1 to 4. Each sentence in the dataset was scored by five evaluators. The final decision for a sentence to be considered as positive was made based on the average value of judgments. We used the conservative threshold according to which a sentence is deemed positive if its average judgement value is over 2.5.

[14] We have also experimented with embedding models but they did not perform better.

[15] `https://www.figure-eight.com/`

[16] We set n=5 as the number of top sentences returned for every top-ranked topic in *Topic-based MRRW*, and for each top-ranked topic pair in *Topic Pair-based MRRW* method and *Topic co-occurrence* methods.

Table 2: Main results.

	P@1	P@5	P@10	P@15	MRR	MAP
Random	5.00	21.00	18.5	18.33	28.81	28.75
Centroid	10.00	18.00	15.00	16.67	28.94	27.10
Topic co-occurrence	15.00	19.00	19.00	20.33	29.58	26.55
MRRW [7]	25.00	28.00	28.00	30.33	46.42	36.94
Topic-based MRRW	**35.00**	27.00	27.00	27.66	**51.54**	39.87
Topic Pair-based MRRW	15.00	**29.00**	**32.00**	**31.33**	50.04	**39.98**

5 Experimental Results

5.1 Main Results

Table 2 shows the overall results according to the Precision@1, 5, 10, 15, Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP).

We found that both of the proposed approaches perform the best on MRR and MAP when compared to the baselines. For the precision, either *Topic-based MRRW* or *Topic Pair-based MRRW* produces the best results depending on the cut-off level. Out of the two proposed approaches, *Topic Pair-based MRRW* appears to be superior, except for P@1 for which *Topic-based MRRW* produces higher quality output. The third best performing method is MRRW, which indicates that graph-based approaches are effective for our task. The satisfactory performance of both proposed approaches, yet with certain differences, suggests also that it may be worthy to experiment with their combination in the future.

Looking at the performance in terms of MRR and MAP over particular categories as shown in Tables 3 and 4, we can observe that although different methods perform best for different category types, the proposed approaches, especially, *Topic-based MRRW* tend to be most stable. The results of the *Topic-based MRRW* have consistently high interestingness rates across all the concept categories. The results for *MRRW* indicate that it has about 8% to 11% drop when compared to the best performing approach, yet it still outperforms the other baselines by a good margin. On the other hand, *Centroid* method, as the most intuitive and simple one, performed quite similar to the *Random* baseline. Similarly, *Topic co-occurrence* – a direct approach that uses a single shared topic space – is not enough to produce effective results.

The *Technology* category seems to be easiest for the interesting content finding task. Most of the tested methods are able to return many interesting contents in this category. This is likely because technology has changed quite much over the last thirty years, and thus facts and opinions from the past on technology-related news are quite different from the present. Technology is ubiquitous these days and perhaps also more appealing to users.

5.2 Case studies

We discuss now a few examples of sentences recommended by our approaches. The first sentence that we want to highlight is the following:

"Of the 715 apartment fires in Moscow last month, 90 were blamed on exploding television sets, a statistic the Soviet press has viewed as an alarming commentary on soviet technology." (Dec 1987)

Table 3: Performance according to different categories by MRR.

	Economy	Places	Politics	Sports	Tech	*Average*
Random	**47.50**	10.49	35.00	23.96	27.08	*28.81*
Centroid	41.67	43.75	10.83	22.62	25.83	*28.94*
Topic co-occurrence	5.20	18.94	8.33	44.58	70.83	*29.58*
MRRW [7]	19.58	**47.92**	25.00	**64.58**	75.00	*46.42*
Topic-based MRRW	40.63	42.36	51.79	58.33	64.58	***51.54***
Topic Pair-based MRRW	43.94	39.40	**52.27**	14.58	**100.00**	*50.04*

Table 4: Performance according to different categories by MAP.

	Economy	Places	Politics	Sports	Tech	*Average*
Random	**38.22**	16.04	**38.57**	23.69	27.25	*28.75*
Centroid	34.38	**39.99**	11.94	17.67	31.50	*27.10*
Topic co-occurrence	6.14	26.67	10.20	31.02	58.71	*26.55*
MRRW [7]	21.93	34.74	19.67	45.37	62.99	*36.94*
Topic-based MRRW	31.07	29.33	32.18	**53.29**	53.47	*39.87*
Topic Pair-based MRRW	34.27	28.40	37.35	15.00	**84.86**	***39.98***

The notion of exploding TV sets in USSR is obviously quite different from our common sense; yet these kinds of unfortunate events were reported several times in 1987[17]. Another example extracted is also rather opposite from what one would claim nowadays:

"Laptop computers are great in theory but disappointing in real life." (Oct 1988)

One could try to explain this example by potentially high expectations put on personal computing tools in the past, coupled with rather low specs of machines at hand and the lack of infrastructure (e.g., wifi spots). Whatever the reasons were, this kind of content might stimulate deliberating about technology evolution and all the "bumps in its evolutionary path" over time. It might serve as an "invitation" for closer reading of the original document or related ones in search for explanation.

Some of the examples from the politics category show certain resemblance to the present day's trade tensions yet the actors are now quite different:

"President Reagan is likely to soon lift some of the trade sanctions imposed on Japan seven months ago during a dispute over Japanese dumping of computer chips, the Administration said today." (Nov 1987)

"Prime Minister Yasuhiro Nakasone today accused the Toshiba Machine Company of betraying Japan by selling militarily sensitive technology to the Soviet Union." (Jul 1987)

[17] Anecdotally, this particular example triggered recollections of childhood memories of one author. His grandparents owned a USSR-produced TV set and often warned him not to sit close to it when he visited their home. Only now, he could understand that the fears of his relatives were actually not without a substance. On a more general note, exploring news archives offers chances for learning about history, and might sometimes even lead to serendipitous discoveries and recollections as this example demonstrates.

We also found opportunities for improvement of our approaches. Take the following two sentences as examples:

"Zenith said the new SX laptop could operate for more than three hours on the battery before it needed recharging." (Oct 1989)

"The Houston-based Company Show edits new battery-operated SLT/286 lap-top system, a computer that it said matches the function of desktop computers but comes in a lunch box-sized, 14-pound package." (Nov 1988)

The news on developments in battery-operated laptops and on battery lifetimes seemed to be frequently reported in the past. However, they do not appear often in the present-day news about laptops. The reason is that battery improvements became rather commonsense nowadays along with the proliferation of producers and, in general, along with the rapid technology progress. Thus they tend not to be special enough to be reported in news articles. Nevertheless, such sentences are returned by our approach (topics popular in the past but not popular now) as our methods do not capture implicit knowledge. Incorporating approaches that use common sense reasoning and analysis as well as extract implicit knowledge could then become advantageous in future research. Another observation based on these examples is that numerical values, such as product specifications (e.g., "14-pound" (or over 6kg) as in the last example), could be extracted and compared to the currently typical ones for finding striking differences. Also, aspects that are obvious at present but were overly emphasized in the past (e.g., "a lunch box-sized" or "battery-operated" as in the above examples) could be considered. Overall, studying elements of surprise and interestingess in archival news could be opening the door for new ideas that lead to automatic approaches for generating/recommending the content of museums and exhibitions.

6 Conclusions

Making document archives more attractive and popular among ordinary users remains a key and perennial goal of the archival community [32,39]. The attractiveness and, related to it, the level of use of document archives among ordinary users is still moderate and can be improved by applying suitable techniques. To this end, we proposed a novel research problem of finding interesting content from news article archives and we approach this challenging task in a fully unsupervised manner. Our key idea is based on data comparisons across time for capturing information surprising to current users. We note that interestingness may have several aspects according to users' age, culture and other backgrounds. The particular, objective measure of interestingness we used in our methods (i.e., surprise arising due to time passage) naturally cannot exhaustively capture the entire spectrum of interestingness.

In the future, we plan to focus on improving the quality of results. As also discussed in [1], it is important to avoid returning trivial and obvious content (in our case, some returned sentences are novel but unsurprising), or one poorly understandable by users, e.g., due to the lack of necessary context.

Acknowledgments. This work has been partially funded by MEXT JSPS Grant-in-Aid. Ricardo Campos, one of the authors of this paper was financed by the ERDF – European Regional Development Fund through the North Portugal Regional Operational Programme (NORTE 2020), under the PORTUGAL 2020 and by National Funds through the Portuguese funding agency, FCT - Fundação para a Ciência e a Tecnologia within project PTDC/CCI-COM/31857/2017 (NORTE-01-0145-FEDER-03185). This funding fits under the research line of the Text2Story project. The first author was employed by Kyoto University when the first version of this paper was created.

References

1. Adamopoulos, P., Tuzhilin, A.: On unexpectedness in recommender systems: Or how to better expect the unexpected. ACM TIST **5**(4), 54 (2015)
2. Baldi, P., Itti, L.: Of bits and wows: A bayesian theory of surprise with applications to attention. Neural Networks **23**(5), 649–666 (2010)
3. Berk, N.A., Gültekin, F.: The topics that students are curious about in the history lesson. Procedia-Social and Behavioral Sciences **15**, 2785–2791 (2011)
4. Berlyne, D.E.: Conflict, arousal, and curiosity. (1960)
5. Boldi, P., Monti, C.: Llamafur: learning latent category matrix to find unexpected relations in wikipedia. In: Proc. of WebScience. pp. 218–222. ACM (2016)
6. Campos, R., Dias, G., Jorge, A.M., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Comput. Surv. **47**(2), 15:1–15:41 (2014)
7. Chen, Y.N., Metze, F.: Two-layer mutually reinforced random walk for improved multi-party meeting summarization. In: Spoken Language Technology Workshop (SLT), 2012 IEEE. pp. 461–466. IEEE (2012)
8. Costa, M., Silva, M.: Understanding the Information Needs of Web Archive Users. The 10th International Web Archiving Workshop (2011)
9. Derezinski, M., Rohanimanesh, K., Hydrie, A.: Discovering surprising documents with context-aware word representations. In: 23rd International Conference on Intelligent User Interfaces. pp. 31–35. ACM (2018)
10. Färber, M.: Semantic Search for Novel Information, vol. 31. IOS Press (2017)
11. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR) **38**(3), 9 (2006)
12. Gomes, D., Cruz, D., Miranda, J., Costa, M., Fontes, S.: Search the past with the portuguese web archive. In: Proceedings of the 22nd International Conference on World Wide Web. pp. 321–324 (2013)
13. Hidi, S., Baird, W.: Interestingness—a neglected variable in discourse processing. Cognitive science **10**(2), 179–194 (1986)
14. Itti, L., Baldi, P.F.: A principled approach to detecting surprising events in video. In: Proc. IEEE Conference on Computer Vision and Pattern Recognition (CVPR). pp. 631–637. San Siego, CA (Jun 2005)
15. Itti, L., Baldi, P.: Bayesian surprise attracts human attention. Vision research **49**(10), 1295–1306 (2009)
16. Kaminskas, M., Bridge, D.: Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. ACM Tran. on Inter. Intell. Systems (TiiS) **7**(1), 1–42 (2016)
17. Kanhabua, N., Anand, A.: Temporal information retrieval. In: Proceedings of the 39th International ACM SIGIR conference on Research and Development in Information Retrieval. pp. 1235–1238 (2016)
18. Koolen, M., Kamps, J.: Searching cultural heritage data: Does structure help expert searchers? In: Adaptivity, Personalization and Fusion of Heterogeneous Information, pp. 152–155. Citeseer (2010)

19. Kuznetsov, S.O., Makhalova, T.: On interestingness measures of formal concepts. Information Sciences **442**, 202–219 (2018)
20. Li, X., Croft, W.B.: Improving novelty detection for general topics using sentence level information patterns. In: Proc. of CIKM. pp. 238–247. ACM (2006)
21. Liu, B., Hsu, W., Mun, L.F., Lee, H.Y.: Finding interesting patterns using user expectations. IEEE Transactions on Knowledge and Data Engineering **11**(6), 817–832 (1999)
22. Macrae, C.N., Bodenhausen, G.V.: Social cognition: Thinking categorically about others. Annual review of psychology **51**(1), 93–120 (2000)
23. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems **27**(3), 303–318 (1999)
24. Pasquali, A., Mangaravite, V., Campos, R., Jorge, A., Jatowt, A.: Interactive system for automatically generating temporal narratives. In: European Conference on Information Retrieval. Springer (2019)
25. Pessent, E.: Is history irrelevant? Dissent Magazine (1971)
26. Sandhaus, E.: The New York Times Annotated Corpus. Linguistic Data Consortium, Philadelphia **6**(12), e26752 (2008)
27. Schwartz, J.M., Cook, T.: Archives, records, and power: the making of modern memory. Archival science **2**(1-2), 1–19 (2002)
28. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE TKDE **8**(6), 970–974 (1996)
29. Silveira, T., Zhang, M., Lin, X., Liu, Y., Ma, S.: How good your recommender system is? a survey on evaluations in recommendation. International Journal of Machine Learning and Cybernetics **10**(5), 813–831 (2019)
30. Silvia, P.J.: What is interesting? exploring the appraisal structure of interest. Emotion **5**(1), 89 (2005)
31. Spyropoulou, E., De Bie, T., Boley, M.: Interesting pattern mining in multi-relational data. Data Mining and Knowledge Discovery **28**(3), 808–849 (2014)
32. Stiller, J.: A framework for classifying interactions in cultural heritage information systems. International Journal of Heritage in the Digital Era **1**(1), 141–146 (2012)
33. Strauss, V.: Why so many students hate history — and what to do about it? The Washington Post (2017)
34. Tran, G., Alrifai, M., Herder, E.: Timeline summarization from relevant headlines. In: European Conference on Information Retrieval. Springer (2015)
35. Trant, J.: Understanding searches of a contemporary art museum catalogue: A preliminary study. Report, Archives & Museum Informatics (2006)
36. Tsukuda, K., Ohshima, H., Yamamoto, M., Iwasaki, H., Tanaka, K.: Discovering unexpected information on the basis of popularity/unpopularity analysis of coordinate objects and their relationships. In: Proc. of SAC. pp. 878–885. ACM (2013)
37. Tsurel, D., Pelleg, D., Guy, I., Shahaf, D.: Fun facts: Automatic trivia fact extraction from wikipedia. In: Proceedings of WSDM. pp. 345–354. ACM (2017)
38. Veale, T., Cardoso, A.: Computational Creativity: The Philosophy and Engineering of Autonomously Creative Systems, vol. 31. Springer (2019)
39. Warwick, C., Terras, M., Huntington, P., Pappa, N.: If you build it will they come? the lairah study: Quantifying the use of online resources in the arts and humanities through statistical analysis of user log data. Literary and linguistic computing **23**(1), 85–102 (2007)
40. Yannakakis, G.N., Liapis, A.: Searching for surprise. In: Proceedings of the International Conference on Computational Creativity (2016)