Towards Recommending Interesting Content in News Archives

I-Chen Hung[1], Michael Färber[12], and Adam Jatowt[1]

[1] Department of Social Informatics, Kyoto University, Japan
[2] University of Freiburg, Germany
ichen@db.soc.i.kyoto-u.ac.jp, michael.faerber@cs.uni-freiburg.de,
adam@dl.kuis.kyoto-u.ac.jp

Abstract. Recently, many archival news article collections have been made available to wide public. However, such collections are typically large, making it difficult for users to find content they would be interested in. Furthermore, archived news articles tend to be perceived by ordinary users as having rather weak attractiveness and being obsolete or uninteresting. In this paper, we propose the task of finding interesting content from news archives and introduce two simple methods for it. Our approach recommends interesting content by comparing the information written in the past with the one from the present.

Keywords: news archive · interestingness · recommender systems

1 Introduction

News article archives offer rich account of events in the past and are important for humanities and social studies constituting one foundation of historical understanding [7]. Besides professionals, news archives can offer valuable content regarding our heritage also for ordinary users. However, non-professional users typically face the following issues when accessing archives. As archives are often of large size, users may quickly get disappointed especially when they do not have any precise search intent. Unstructured content and the typically unknown context of the past can easily cause confusion and boringness. Finally, the content may seem obsolete and detached from the present.

We believe that special kind of information recommendation for large size news archives could increase their utility and attractiveness for average users. The recommended information should be related to present issues yet not be obvious or inferable, and should preferably contain an element of surprise. Let us consider "ice cutter" as an example. Ordinary users might nowadays expect that *ice cutter* was a machine that cut ice into small pieces. In the past, however, especially before the widespread use of refrigerators, *ice cutter* was a person who cut ice from frozen lakes and rivers. Such content has potential to surprise non-professional readers and evoke their interest, as the contained information is against the presumed expectation. Note that such information is not easy to be found using traditional search engines.

We propose in this paper a novel research problem which, to the best of our knowledge, has not been pursued so far: *interesting content recommendation from long-term news archives*. One of the main difficulty in finding interesting pattern or data is how to define *interestingness* properly. Geng *et al.* [4] treated interestingness as a broad concept that possibly contains features like diversity, surprisingness and so on. Silberschatz *et al.* [8] suggested that interesting information should be unexpected and actionable. Yannakakis *et al.* [10] proposed a surprise-oriented search algorithm. Unexpectedness was also considered crucial in Padmanabhan *et al.* [6] and Adamopoulos *et al.* [1]. Although there have been a few studies about how to identify content about the unexpected relationships, they focused on non-archival data such as Wikipedia or on current news [5]. Tsukuda *et al.* [9] evaluated the unexpectedness of related terms extracted from Wikipedia pages on the basis of relationships of the coordinate terms. Boldi *et al.* [2] focused on finding unexpected links in hyper-linked documents. None of the prior works however focused on archival content which has particular characteristics due to time passage.

In this paper, we propose four criteria of content interestingness in news archives: (1) *Relevance*: interesting past content should be relevant to user query and (2) *Past importance*: it should be important (not minor) in the past. (3) *Unfamiliarity*: interesting past content should be unknown to a user and (4) *Unexpectedness*: it should be unexpected or surprising to her.

In our approach we assume that users can input general queries representing their interests, and we output ranked lists of sentences considering some of the proposed criteria of archival content interestingness.

2 Approach

In our approach we divide the underlying news archive into two parts: one denoted as D_{past} and representing past documents (i.e., documents published at some period T_{past} in the past), and the other one, denoted as D_{now}, that contains documents published recently (i.e., in a recent period T_{now}) to represent information of present. Sentences from D_{past} that are relevant to user query will be ranked based on their comparison with relevant sentences in D_{now}. We propose two methods as follows:

Centroid method. We hypothesized that interesting content should be unfamiliar and unexpected to current users. Centroid method will then rank sentences from D_{past} by their dissimilarity to the centroid vector being the average TF-IDF vector of all sentences in D_{now}. Centroid method is expected to extract relevant sentences that have less chance to be known by current users.

MRRW. The *two-layer mutually reinforced random walk* (MRRW) [3] is algorithm for computing the converged scores of nodes in a two-layer graph. We adapt the algorithm considering two time periods: T_{past} and T_{now} such that nodes in each layer represent relevant sentences in the corresponding document set. For both layers T_{now} and T_{past}, we connect nodes belonging to the same layer by calculating their similarity (cosine similarity of sentences represented by the

nodes). On the other hand, a node pair consisting of nodes from different layers is connected by an edge whose weight represents the nodes' dissimilarity. MRRW process will return sentences from D_{past} that are similar to other sentences in T_{past}, yet, at the same time, are dissimilar to sentences in D_{now}. This approach is expected to reflect the *past importance* and *unfamiliarity* apart from *relevance*, since a sentence dissimilar to many sentences in T_{now} might represent novel and unfamiliar content for users.

3 Experiments

3.1 Datasets and Experimental Setup

We use the New York Times News Archive[3] which includes news articles published from 1987 to 2007. In our experiments, D_{past} contains articles published from 1987 to 1989 and D_{now} covers ones published from 2005 to 2007. Naturally, the latter part is not exactly representing the "present", and is a compromise resulting from the lack of sufficiently long datasets which would contain also recent documents. We assume that all sentences containing a query are relevant to it. We use TF-IDF and Word2vec embeddings trained on entire archive for sentence representation, and conduct experiments using 20 different queries covering equally topics from economy, politics, sports, technology and the names of geographic locations. We test **Centroid**, **TF-IDF+MRRW** and **Word2vec+MRRW** methods as well as a baseline approach based on random sentence selection **Random**.

We let 15 evaluators (6 males and 9 females in their 20s and 30s with at least bachelor level education) judge the quality of sentences. In particular, they were asked to assess the results based on *Understandability* and *Interestingness*. Evaluators needed to label the sentences by either *yes* or *no*. Before the evaluation, we extracted and pooled the top 15 returned results for each query by each tested method. Each sentence was evaluated by three evaluators, and only sentences understood by evaluators (as given by a binary *understandability* score) were considered valid. The final decision if a sentence is interesting was made based on the majority vote, i.e., if at least two evaluators gave it a score of 1.

3.2 Experimental Results

Table 1 shows the results according to Mean Reciprocal Rank (MRR) and Mean Average Precision (MAP) for 5 query categories. We could see that **Centroid** method performs quite well but the performance of methods vary between query categories. This suggests that different query categories have different best performing methods and thus one should be careful when choosing suitable recommendation approach for a given query.

Random baseline has notably high precision only on *Technology* queries. Despite of weaker performance of **Word2vec+MRRW**, it still achieved high

[3] http://www.nytimes.com/ref/membercenter/nytarchive.html

Table 1. Performance of methods for each query category.

	MRR					MAP				
	Economy	Places	Politics	Sports	Tech	Economy	Places	Politics	Sports	Tech
Random	70.83	48.96	23.75	44.20	**70.83**	52.90	43.24	24.17	36.67	**72.68**
Centroid	**75.00**	**87.50**	**33.33**	53.57	66.67	**65.71**	**71.92**	**29.83**	53.29	59.80
TF-IDF+MRRW	70.83	**87.50**	23.61	**70.83**	10.42	51.12	69.88	28.19	**62.40**	9.51
Word2vec+MRRW	31.25	81.25	31.25	**70.83**	27.08	38.44	66.09	28.50	46.19	31.41

scores on *Sports* and *Places*. One possible improvement could be inputting more data for model training.

4 Conclusions & Future Work

In this paper, we propose a novel research problem of recommending interesting contents from news article archives and we describe our initial approach. Our key idea is based on data comparison across time. In future, we plan to improve the quality of results to avoid outputting trivial content or one poorly understandable due to the lack of necessary context. We will also consider other interestingness criteria that we did not explicitly include this time in our approach and time period suggestion from which interesting results can be derived.

Acknowledgments. This research was supported by MEXT grants (#17H01828; #18K19841; #18H03243).

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. Boldi, P., Monti, C.: Llamafur: learning latent category matrix to find unexpected relations in wikipedia. In: Proceedings of WebScience. pp. 218–222. ACM (2016)
3. 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)
4. Geng, L., Hamilton, H.J.: Interestingness measures for data mining: A survey. ACM Computing Surveys (CSUR) **38**(3), 9 (2006)
5. Li, X., Croft, W.B.: Improving novelty detection for general topics using sentence level information patterns. In: Proceedings of the CIKM. pp. 238–247. ACM (2006)
6. Padmanabhan, B., Tuzhilin, A.: Unexpectedness as a measure of interestingness in knowledge discovery. Decision Support Systems **27**(3), 303–318 (1999)
7. Schwartz, J.M., Cook, T.: Archives, records, and power: the making of modern memory. Archival science **2**(1-2), 1–19 (2002)
8. Silberschatz, A., Tuzhilin, A.: What makes patterns interesting in knowledge discovery systems. IEEE TKDE **8**(6), 970–974 (1996)
9. 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)
10. Yannakakis, G.N., Liapis, A.: Searching for surprise. In: Proceedings of the International Conference on Computational Creativity (2016)