Finding and Recommending Interesting Contents from Document Archives

I-Chen HUNG†, Michael FÄRBER†, and Adam JATOWT†

† Graduate school of Informatics, Kyoto University
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, JAPAN
E-mail: †ichen@db.soc.i.kyoto-u.ac.jp, ††michael.faerber@cs.uni-freiburg.de, †††adam@dl.kuis.kyoto-u.ac.jp

Abstract In recent years, many archival collections have been digitized and made available on Web. However, archives are typically of large size, and it is difficult for users to find the content they are interested in. One problem regarding the content stored in archives, which results from the inherent characteristic of document archives, is its perceived weak attractiveness for current users and relevance to the present issues. Past archival documents may seem detached from the present and appear obsolete; however, at the same time, this can also mean the archives can become interesting if suitable content can be found and extracted. In this paper, we propose a model for finding interesting content from long-term document archives and we design a method for outputting such content given a user query. In particular, we take the New York Times news corpus as data input, and extract contents based on comparisons between two different time periods. Using the proposed method and basing the approach on the findings from the psychology and cognitive sciences, we aim to create a system that can generate the interesting output from the input.

Key words archive; interestingness; information recommendation

1. Introduction

An archive is an accumulation of historical records that have been considered as social constructs [15]. The need of use changes by time, and the size of data stored in archives tend to grow larger and larger, which may result in two disadvantages for current users interested in finding information from past data：

(1) It is a heavy burden to read all of the contents, which is also a very inefficient task. Document archives are often stored as raw text format after being digitalized. The property of unstructured content and the unknown context of the past easily causes confusion and boringness. Also, it is very difficult to find the information users are interested in among the numerous text data.

(2) Users will possibly get numerous and unattractive results if they use traditional information retrieval methods on the document archives. Most of the search methods return the search results regarding the relevance and popularity to the input queries. However, due to the inherent characteristic of document archives, the document archives are often considered less attractive and seem detached from present issues.

Therefore, information recommendation necessary for large size document archives to increase their utility and attractiveness for average users. In order to solve the above-mentioned shortcomings, the recommended information from document archives should be interesting for current users. To be more specific, an interesting information should be related to present issue yet not obvious or inferable. Lets take "ice cutter" as an example. Users would likely expect it to mean:

> "Ice cutter is a kind of machine that can cut huge ice into small pieces."

However, two hundred years ago, the meaning of "ice cutter" was as follows:

> "Ice cutter was a job. Before widespread use of refrigerator, ice was cut from frozen lakes and rivers by men."

The later content snippet demonstrates potentially interesting content to document archive users as it is rather against the presumed expectation of current users. Naturally, such information could be found using a search engine. However, it requires some efforts and search skills for users to find interesting content on the past in the current Web. In order to make search and access to past archives interesting to nowadays' users special kinds of information recommendation such as recommendation of unexpected content should be helpful. Although there have been a few studies about

how to identify the unexpected relationships, they focused on non-archival data, such as Wikipedia [4] [18] or current news [11]. As mentioned above, contents in archives are often stored as raw text format, thus the existing approaches seems to be not appropriate for our case.

In this paper, we focus on extracting sentences based on the objective measures of interestingness. We propose four major attributes of archival content which are important to consider the content interesting from the perspective of psychology and cognitive science:

- Relevant to query
- Not minor in the past
- Novel and unfamiliar to user
- Unexpected and surprising

We aim to extract and recommend sentences from documents stored in archives based on the input query and considering the proposed metrics of interestingness, which we will discuss further in Section 3. We adapt the two-layer mutually reinforced random walk to capture the novelty and unexpectedness in archives among time periods. Our experiments are performed on New York Times news corpus from 1987 to 2007. There are two main objectives of the research:

(1) To show the attractiveness of long-term document archives to current users. Due to the change of time, users might consider the content in the archives less related to current issues and less attractive. With the recommendation methods that center on interestingness, we are able to show users information that they feel unexpected, new, and might be potentially interested in.

(2) To define interestingness from psychological and cognitive science perspectives and to confirm it by a computational method. Being interesting is an abstract concept of the emotional status such that people's attention have been captured and their curiosity have been aroused. However, the previous studies have no clear consensus on the measures of interestingness [8].

The rest of the paper is organized as follows. Section 2 gives a brief review of related works. Section 3 introduces the definitions of interestingness used in this research, which are supported by psychological viewpoints. Section 4 proposes methods for discovering the interesting patterns in archives according to the input query. Section 5 describes the experiment set-up and shows the results in using the proposed approach. A summary of this research and the future works are given in Section 6.

2. Related Works

Our work is related to several research topics as follows:

Definition of interestingness. One of the main problem in finding interesting pattern or data is how to define

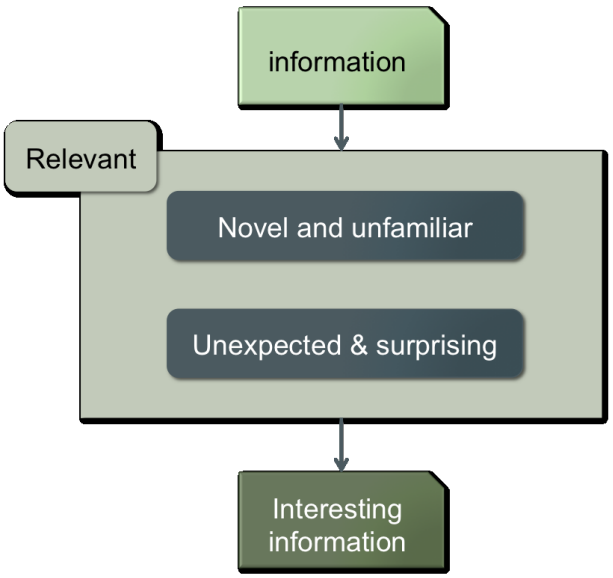

Figure 1 The overview of the interestingness based recommendation system

interestingness properly. Geng *et al.* [8] treat interestingness as a broad concept that possibly contains features like reliability, diversity, surprisingness and so on. In the task of pattern finding in knowledge discovery system, Silberschatz *et al.* [16] focus on the subjective measures of interestingness, suggesting that interestingness should be unexpected and actionable. Unexpectedness is also considered crucial in Padmanabhan *et al.* [13] and Adamopoulos *et al.* [1]; moreover, the latter one introduce serendipity as one of their evaluation measures.

Unexpected relationship detection. There are several studies that focus on the approach of finding the unexpected relationships. Both Boldi *et al.* [4] and Tsukuda *et al.* [18] use Wikipedia(注1) as their experiment datasets. Tsukuda *et al.* [18] evaluate the unexpectedness of related terms extracted from Wikipedia page on the basis of relationships of the coordinate terms. Boldi *et al.* [4] focus on finding unexpected links in hyperlinked document.

On the other hand, Adamopoulos *et al.* [1] calculate the unexpectedness score with user expectations in the recommendation system, and take utility of results into account. And Jacquenet *et al.* [9] take the document structure into account, using four representations to do the similarity detection.

Novelty detection. TREC challenge(注2) consists of a a set of tracks and tasks, such as TREC Temporal Summarization (TempSum), TREC Knowledge Base Acceleration (KBA), and TREC Novelty Track, and has brought the improvement of the novelty detection for years Farber *et al.* [7].

(注1)：https://www.wikipedia.org/
(注2)：http://trec.nist.gov/

Features like *sentence lengths*, *name entities*, and *opinion pattern* are used in Li *et al.* [11] to analyze and improve the novelty detection on the 2002-2004 TREC novelty tracks.

3. Definition of Interestingness

We believe that the definition of interestingness is very important for the performance of our archive-based content recommendation approach.

3.1 Relevant to Query and not Minor in the Past

First of all, the target information should be relevant to the query user input. Yet, considering the possible change of time and issues presented in news, we target content related to the input query yet not trivial and minor in the past.

3.2 Novel and Unfamiliar to User

Sequential check theory of emotion differentiation is described as a part of a dynamic model of emotion process in [14]. The theory explained the emotional state as the result of a series stimulus evaluation, which also called appraisal check, and make prediction to the oncoming response. *Novelty check* is the at the primary level of relevance detection, which is the one of four major type of appraisal objectives.

When people perceived a sudden stimulus from outside, in order to decide to put attention to the new stimulus or nor, the relevance check mechanism will be evoked and then determine the degree of familiarity with the object event. In Silberschatz *et al.* [16], interestingness is considered as an emotion that have clear motivational and goal component for exploring and learning, which means the objective of interestingness is on something unfamiliar and new.

In this metric, we target the information that is novel and unfamiliar to current users from the past document archives, which we believe would effectively arouse users' attention and curiosity.

3.3 Unexpected and surprising

Since the limitations of the basic cognitive and the overloading stimulus from environment, it is necessary for people to simplify the perception process. According to Macrae *et al.* [12], this goal can be achieved through categorical thinking. Perceiver tends to observe event on the basis of social categories, such as gender and age. Therefore, it is easy to identify unexpected stimulus from expected one by checking the properties with the social categories it belongs to.

To be more specific, the unexpected and surprising information is that does not conform to available stereotypical expectation. For example, an elder in fashionable coat and with tattoo might be very impressive due to the inconsistency between the stereotype of ordinary elders and the specific case. In this metric, we target the information that is different from the obvious common content.

To sum up, we are going to propose a method to extract the interesting information that has been filtered by these metrics.

4. Architecture of the Recommendation system Considering Interestingness

In this section, we will briefly introduce the process of how to find and recommend interesting information.

4.1 Input

The input to the recommendation system are two document archives and a set of queries. The two document archives, $D_{now} = \langle d_1, d_2, d_3, ..., d_i \rangle$ represent the information from T_{now}, and the other one $D_{past} = \langle d_1, d_2, d_3, ..., d_j \rangle$ represented the information from T_{past}, contain content of general topic. Besides queries, users could also specify the time span for D_{past}. D_{now} will be set to the latest dataset in the document archives by default. For example, $q = \{ice, cutter, food\}, T_{past} = [1987, 1989]$ will process the searching and recommendation based on the q covering the time period from 1987 to 1989 of the target document archives.

4.2 Preprocessing for Interestingness Measures

Our goal is to find the relationship and similarity between all the target documents. In order to capture the second and third proposed metrics, *Novel and unfamiliar to users* and *Unexpected and surprising*, mentioned in the previous section, we adapted the two-layer mutually reinforced random walk (MRRW) algorithm [5]. The MRRW algorithm will finally return the information that is similar to information within the same layer yet dissimilar to those in the different layer. We will give detailed explanation in section 4.3.

Firstly, we trim the document contents by removing stopwords and punctuations. In order to fit in the random walk model, we tokenize both D_{now} and D_{past}, fitting into TF-IDF model and calculating the cosine similarity.

The process will be done for three times in total: One for cosine similarity calculation between the nodes within T_{past} layer, in which a node is a TF-IDF vector representing a document d_j in D_{past}; one for cosine similarity calculation between the nodes within T_{now} layer; and the other for dissimilarity calculation between nodes in different layers.

4.3 Interestingness Detection by MRRW

For layer T_{now} and T_{past}, we connect each node pair belonging to the same layer by calculating the node similarity, and each node pair belonging to different layer by calculating the node dissimilarity.

Let denote layer T_{past} as $L_{PP} = \{n_{d1}, n_{d2}, ..., n_{di}\}$, and layer T_{now} as $L_{NN} = \{m_{d1}, m_{d2}, ..., m_{dj}\}$, where n_{di} and m_{dj} are TF-IDF vector for a document. The edges within layer are computed as:

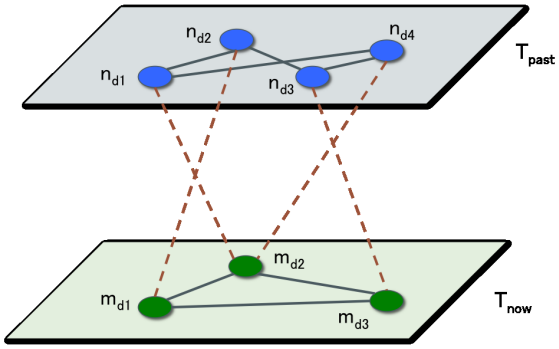

Figure 2 The overview of two-layer mutually reinforced random walk (MRRW). n_{di} represents the TF-IDF vector of document d_i, and m_{dj} represents the TD-IDF vector of document d_j. The edge describes the similarity of the two connected nodes within L_{PP} and L_{NN}, and the dotted line edge describes the dissimilarity of nodes between the layers.

$$Sim(n_{di}, n_{di}) = \frac{\overrightarrow{n_{di}} \cdot \overrightarrow{n_{di}}}{|\overrightarrow{n_{di}}| \times |\overrightarrow{n_{di}}|} \quad (1)$$

$$Sim(m_{dj}, m_{dj}) = \frac{\overrightarrow{n_{dj}} \cdot \overrightarrow{n_{dj}}}{|\overrightarrow{n_{dj}}| \times |\overrightarrow{n_{dj}}|} \quad (2)$$

The edge weight between layer is computed as:

$$DisSim(n_{di}, m_{dj}) = 1 - \frac{\overrightarrow{n_{di}} \cdot \overrightarrow{m_{dj}}}{|\overrightarrow{n_{di}}| \times |\overrightarrow{m_{dj}}|} \quad (3)$$

For finding the information which is relatively unfamiliar to current users yet at the same time not trivial in the past, we adapt the two-layer mutually reinforced random walk (MRRW) [5] to reinforce the score for each node.

$$\begin{cases} S_P = (1-\alpha)S_P + \alpha \cdot L_{PP}L_{PN}S_N \\ S_N = (1-\alpha)S_N + \alpha \cdot L_{NN}L_{NP}S_P \end{cases} \quad (4)$$

Here S_P and S_N denote the interestingness score in L_{PP} and L_{NN} respectively. As Figure 2 shows, after we apply Eq. (4) to our datasets, the score will become higher if the node is more similar to the nodes in the same layer, but dissimilar to the nodes in the other layer. For example, the score of n_{d1} will be highly reinforced if $Sim(n_{d1}, n_{d2})$, $Sim(n_{d1}, n_{d4})$ and $DisSim(n_{d1}, m_{d2})$ are all high, such as 0.9, 0.8 and 0.85. On the other hand, if $Sim(n_{d2}, n_{d3})$ and $Sim(n_{d3}, n_{d4})$ are high, yet $DisSim(n_{d3}, m_{d3})$ is row, such as 0.3, the final score of n_{d3} will be affected by the row $DisSim$ score and thus become lower.

4.4 Expected Output

The expected output will be the most relevant and unexpected sentence which is determined by the final score reinforced by the proposed model between and within L_{PP} and L_{NN}. For considering the comprehensibility, we will also present the result in paragraph or in document.

5. Experiment

5.1 Datasets

In this research, we use the New York Times News archive(注3) for our experiments. To be more specific, the corpus includes news archives during 1987 to 2007 stored in xml format, containing meta data labels such as date, title, category, full-text, and so on. In the experiments, we use archives from 1987 to 1989 as information of T_{past}, which is denoted as D_{past}, and archives from 2005 to 2007 as information of T_{past}, which is denoted as D_{past}.

5.2 Results

We use $q = laptop$, $q = digital$ as our demo result of the proposed system. Some of the results are presented as follows:

Table 1 The result of query *laptop* from news archives in T_{past}

Score	File ID	Extracted content
0.005034	37717	The Data General Corporation, based in Westboro, Mass., has introduced a new version of its Model One laptop personal computer, which features a new screen, a faster microprocessor, a built-in hard disk and internal removable batteries.
0.004462	291448	In a development that will make small "notebook" computers that weigh four to seven pounds comparable in performance to many desktop machines, the Compaq Computer Corporation plans to introduce on Monday two laptop computers.
0.004454	32568	But in laptop computers, the Japanese models made by Toshiba and NEC are among the most popular.
0.004354	190603	A battery-powered laptop computer was finally introduced last week by the Compaq Computer Corporation.
0.003342	189068	The Houston-based company showed its new battery-operated SLT/286 laptop system, a computer that it said matches the function of desktop computers but comes in a lunch box-sized, 14-pound package.

The score represents the interestingness score S_P of query *laptop*. As we can see in Table 1, the query returned some news about the new released laptops. From the results, we could find out that the Japanese brands were popular choice then. Also, weight seems to be an important element in the past, yet it has become too common to mention in present

(注3)：http://www.nytimes.com/ref/membercenter/nytarchive.html

time.

Table 2 shows the results of *technology* in T_{past}, which contain some statements of technology competition during Cold War. Large part of the extracted news focused on military technology and we could also find out that many countries are still trying to accelerate the technology development.

Table 2 The result of query *technology* from news archives in T_{past}

Score	File ID	Extracted content
0.000152	107712	But United States technology is still better than Soviet technology, which is one of the reasons they seek our support.
0.000145	121357	For the first time, the U.S. would no longer dominate the critical technologies needed for military power and industrial development
0.000138	22532	The new corporation will work with research institutes to accelerate the development of China's radio and television industry, strengthen technological cooperation with foreign companies and help Chinese enterprises to import new technology, key equipment and component parts, the report said.
0.000130	83598	A French company improperly exported advanced American chip-making equipment to the Soviet Union, in a deal that American officials said today would improve Soviet military technology.
0.000130	177342	Many experts contend that even without direct Japanese investment in American high technology, it would be virtually impossible today to keep new computer architectures, like the kinds of computers now made by MIPS and Ardent, exclusively in one nation for long.

6. Conclusion

Many historical archives have been digitalized and can be easily accessed nowadays. Yet many of the valuable contents are overlooked by users due to the large size of text and the presumed lack of their connection to present issues. In order to satisfy the searching need, and decrease cognitive burden on the users when they are trying to find information related to queries, we proposed a method to improve the searching result by using interestingness measures, which are (1) Relevance to query and being not minor in the past, (2) Novelty and unfamiliarity to user, and (3) Unexpectedness and surprise.

In this research, we take the New York Times news corpus as data input, and extract interesting contents in the past based on comparisons between two different time periods. Using the proposed method and basing the approach on the findings from the psychology and cognitive sciences, we aim to create a system that can generate the interesting output from the input.

In the future, we plan to handle the quality concern of results, which has also been discussed in Adamopoulos *et al.* [1]. It is important to avoid trivial and obvious content especially when our searching process is regarding several interestingness metrics. We also consider to take Coping potential Berlyne *et al.* [3] as an additional metric of interestingness based on psychology for improving our experiment result.

7. Acknowledgments

This research and development work was supported by the MIC/SCOPE #171507010.

References

[1] Panagiotis Adamopoulos and Alexander Tuzhilin. On unexpectedness in recommender systems: Or how to better expect the unexpected. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 5(4):54, 2015.

[2] Ching-man Au Yeung and Adam Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1231–1240. ACM, 2011.

[3] Daniel E Berlyne. Conflict, arousal, and curiosity. 1960.

[4] Paolo Boldi and Corrado Monti. Llamafur: learning latent category matrix to find unexpected relations in wikipedia. In *Proceedings of the 8th acm conference on web science*, pages 218–222. ACM, 2016.

[5] Yun-Nung Chen and Florian Metze. Two-layer mutually reinforced random walk for improved multi-party meeting summarization. In *Spoken Language Technology Workshop (SLT), 2012 IEEE*, pages 461–466. IEEE, 2012.

[6] Marie-Catherine De Marneffe, Anna N Rafferty, and Christopher D Manning. Finding contradictions in text. In *ACL*, volume 8, pages 1039–1047, 2008.

[7] M Färber. *Semantic Search for Novel Information*, volume 31. IOS Press, 2017.

[8] Liqiang Geng and Howard J Hamilton. Interestingness measures for data mining: A survey. *ACM Computing Surveys (CSUR)*, 38(3):9, 2006.

[9] François Jacquenet and Christine Largeron. Discovering unexpected documents in corpora. *Knowledge-Based Systems*, 22(6):421–429, 2009.

[10] Adam Jatowt and Marc Bron. Historycomparator: Interactive across-time comparison in document archives. In *COLING (Demos)*, pages 84–88, 2016.

[11] Xiaoyan Li and W Bruce Croft. Improving novelty detection for general topics using sentence level information patterns. In *Proceedings of the 15th ACM international con-*

ference on Information and knowledge management, pages 238–247. ACM, 2006.

[12] C Neil Macrae and Galen V Bodenhausen. Social cognition: Thinking categorically about others. *Annual review of psychology*, 51(1):93–120, 2000.

[13] Balaji Padmanabhan and Alexander Tuzhilin. Unexpectedness as a measure of interestingness in knowledge discovery. *Decision Support Systems*, 27(3):303–318, 1999.

[14] Klaus R Scherer. Appraisal considered as a process of multilevel sequential checking. *Appraisal processes in emotion: Theory, methods, research*, 92(120):57, 2001.

[15] Joan M Schwartz and Terry Cook. Archives, records, and power: the making of modern memory. *Archival science*, 2(1-2):1–19, 2002.

[16] Abraham Silberschatz and Alexander Tuzhilin. What makes patterns interesting in knowledge discovery systems. *IEEE Transactions on Knowledge and data engineering*, 8(6):970–974, 1996.

[17] Paul J Silvia. What is interesting? exploring the appraisal structure of interest. *Emotion*, 5(1):89, 2005.

[18] Kosetsu Tsukuda, Hiroaki Ohshima, Mitsuo Yamamoto, Hirotoshi Iwasaki, and Katsumi Tanaka. Discovering unexpected information on the basis of popularity/unpopularity analysis of coordinate objects and their relationships. In *Proceedings of the 28th Annual ACM Symposium on Applied Computing*, pages 878–885. ACM, 2013.