Discovering Latent Threads in Entity Histories

Yijun Duan¹ · Adam Jatowt¹ · Katsumi Tanaka¹

Received: 20 August 2019 / Revised: 18 October 2019 / Accepted: 22 October 2019 / Published online: 15 November 2019
© The Author(s) 2019

Abstract

Knowledge of entity histories is often necessary for comprehensive understanding and characterization of entities. Yet, the analysis of an entity's history is often most meaningful when carried out in comparison with the histories of other entities. In this paper, we describe a novel task of *history-based entity categorization* and *comparison*. Based on a set of entity-related documents which are assumed as an input, we determine latent entity categories whose members share similar histories; hence, we are effectively grouping entities based on the correspondences in their historical developments. Next, we generate comparative timelines for each determined group allowing users to elucidate similarities and differences in the histories of entities. We evaluate our approach on several datasets of different entity types demonstrating its effectiveness against competitive baselines.

Keywords Entity categorization · Entity comparison · Digital history

1 Introduction

Grouping is a common procedure used for organizing, structuring and understanding entity sets. For example, Wikipedia, which is considered the most comprehensive encyclopedia, contains over 1.13 million categories [2] which consist of multiple members that share some common traits (e.g., list of cities in Japan, list of American Nobel Prize winners, etc.). However, few existing entity groups and categories are based on the common historical patterns of the entities which belong to them. Rather, entities tend to be grouped solely based on their “present” attributes (e.g., cities in the same country, professionals doing similar jobs, car models having similar shapes or parameters), while their histories are rarely used as the sole basis for grouping. Yet, historical aspects are often quite important and, in many cases, the history shapes and define the present characteristics of entities. In fact, quite many entities cannot be fully understood

without the thorough knowledge of their histories. For example, to know a given city well, one definitely needs to study its history because the shape, the atmosphere, the character and most of other “present” attributes of the city are the result of particular historical developments in the past. Similarly, for the case of persons, to understand someone's career or achievements, we try to study the early events in the person's early life and their impact on the subsequent career. Moreover, the histories of certain class of entities directly determine their perceived values (e.g., historical buildings in touristic cities, antiques, precious stamps). The historical accounts related to those kind of objects are strongly associated with their attractiveness and originality. We thus think that constructing history-centered entity groupings can prove beneficial in the case of analyzing certain types of entities, and it can lead to novel and useful ways of entity understanding and comparison.

How could cities in Asia be grouped according to their historical similarities? What are common patterns according to which electronic companies were developing in Japan? What are the different types of biographies of German scientists in the twentieth century? Questions of this type are not easy to be answered because analyzing the histories of all the members belonging to a given input entity group is quite an arduous task, not to say anything about structuring of such a group (e.g., in Wikipedia, there are over 530 articles of major cities in Japan alone, where each

✉ Yijun Duan
minsak1020@gmail.com

Adam Jatowt
adam@dl.kuis.kyoto-u.ac.jp

Katsumi Tanaka
tanaka.katsumi.85e@st.kyoto-u.ac.jp

¹ Graduate School of Informatics, Kyoto University,
Yoshida-Honmachi, Sakyo-ku, Kyoto 606-8501, Japan

Fig. 1 Typical histories of three latent history-based categories of Japanese cities which were learned from 532 instances using our proposed system. Each category is represented in two ways: by a typical instance city (which we call an exemplar), as well as by a timeline consisting of major events ordered chronologically by their computed median occurrence time (which we call a prototype). Each event clus-

ter in the prototype is illustrated by a manually created label positioned on timeline along with the median (left value) and the standard deviation of the occurrence time (right value) of event instances falling into the cluster. As we can notice, the learned categories exhibit diverse patterns of city development, which represent a novel kind of historical knowledge generation and organization

usually has a long dedicated section describing its history). Note that currently entities are usually not grouped based on their common histories. For example, there are no dedicated list pages in Wikipedia for subgroups of Japanese cities or American scientists that would share similar historical developments.

To enable history-based entity grouping, we formulate the *latent history-based category hypothesis*, which states that *entities can be categorized based on the similarity of their histories, such that entities included in the same category have more similar histories to each other than to ones in other categories*. Figure 1 shows the example of the three latent history-based categories that our approach identified out of the input collection composed of the histories of multiple Japanese cities. In order to group entities based on their histories, we devise a concise optimization model (OM) approach inspired by the popular affinity propagation (AP) algorithm for exemplar-based clustering. A number of successful optimization formulations have been developed for diverse natural language processing tasks such as semantic role labeling [28] and document summarization [38]. However, to the best of our knowledge, except for our

work, none has proposed optimization formulation which would aim at *history-driven categorization*.

Since history-based categories can be too large to allow users easily understand common aspects of its members, we additionally propose to describe each generated category. Cognitive science studies suggest two modes in which people can effectively understand and memorize categories: *exemplar theory* [6] and *prototype theory* [27]. Both modes embody the idea of *graded structure* of a category, according to which some members of a category are more central than others. *Exemplar theory* posits the usage of real member entities as exemplars to represent the category, while an alternative *prototype theory* proposes describing category by constructing the “average” of all members as an abstract prototype. Entities more similar to the exemplar or to the prototype are considered better examples of the category. We then first propose to represent each generated latent category by the *exemplar* toward a more efficient understanding. The advantage of using *exemplars* for category explanation is that they represent the actual member of the category and are easy to be remembered [15, 22]. What is more, good exemplars are often more effective to describe shared aspects in history than the descriptions of separate features (in our

case events) because of the high coherence of the former. Still however, the full comprehension of category's history cannot be done using only a single selected entity. This is because a selected exemplar may not embody all the key shared aspects of the category (e.g., we might not be able to pick up just one city that would adequately reflect historical developments of the entire group of cities that have similar histories). We then also generate a summarized timeline (i.e., *prototype*) for each category indicating the similarities of its members. Since we form several latent groups of entities, the creation of the prototype should be also carried out based on the comparison to the histories of other related categories. Thus, besides the common historical similarities of its category, the exemplar will also reflect the dissimilarities to other categories for enabling effective characterization and comparison. With such a dual representation of each latent category (exemplar and prototype), we believe that users should be able to more easily understand the common and distinguishing historical aspects of the derived history-based categories. The advantages of the exemplar are that it is an actual entity, whose history is coherent and real. On the other hand, prototype can better capture common traits in the histories of entities belonging to a given group. The disadvantage here can be, however, weak coherence between the events listed on the timeline and the fact that the presented history is not real in a sense that it is the result of agglomerating histories of multiple entities.

The task of history-focused entity categorization and comparison assumes taking an input set of entities (typically a large category of entities, such as cities, celebrities or scientists from a given country and so on) along with their historical descriptions and then dividing them into latent sub-categories based on historical similarities. Each formed sub-category is then represented by a selected exemplar and a constructed prototype which are supposed to reflect shared historical traits of the members of each sub-category and the differences to the histories of other sub-categories. The task is challenging due to the following reasons.

1. First of all, the proposed novel categorization style should be based on similarity between entities' histories instead of on straightforward text similarity as typically adopted in entity clustering.
2. Grouping entities not only should be based on similarity of their historical events but should also consider the effect of event time as well as incorporate the notion of event importance (i.e., less important or trivial events should play smaller roles in the grouping process). It is, however, not obvious how to automatically estimate event importance.
3. Identifying the number of latent groups is not trivial.

4. Lastly, selecting good exemplars among entities and constructing representative and discriminative prototype timelines are challenging.

Creating the novel kinds of categories that are based on entity histories can be helpful in many scenarios. For instance, history-focused categories could be used for enriching Wikipedia by adding generated latent timelines into Wikipedia list pages (see Fig. 1 for the examples of generated timelines). From a more general viewpoint, users could benefit from the constructed historical knowledge for supporting various kinds of analyses including evolution and causality analysis, finding historical explanations, provenance investigation and for answering history-related questions.

To sum up, we make the following contributions in this paper:

1. We describe a new research problem of *automatically discovering history-based categories* and *constructing their comparative timelines*.
2. We propose to represent the constructed categories by two means, by *exemplar* which is the most representative entity belonging to each category and by *prototype* which is a compact timeline summarizing category history.
3. We develop an unsupervised approach for these tasks based on a concise optimization model and mutually reinforced random walks.
4. The effectiveness of our methods is proved by experiments on seven datasets and by comparison with competitive baselines.

The remainder of this paper is organized as follows. The related works are introduced in Sect. 2. We formulate our research problem in Sect. 3. Section 4 introduces the approach for event importance calculation. The proposed optimization formulation for latent history-based entity categorization is presented in Sect. 5. Section 6 discusses the creation of comparative timelines for all groups. In Sect. 7, we describe the experiments and evaluation results. Section 8 details our discussion on several relevant aspects to our task. Finally, we draw conclusions and outline future work in Sect. 9.

2 Related Work

To the best of our knowledge, besides our work, the research problem of forming history-based entity groups has not been proposed neither approached so far by others. Our work is nevertheless connected to several research areas: *exemplar*

detection for entity categorization and different settings of *document summarization*, that will be described below.

2.1 Exemplar Detection

Cognitive science describes the way in which people understand and memorize categories using the *exemplar theory* [6], which states that individuals make category judgments by comparing new stimuli with instances already stored in their memory. The instance stored in memory is regarded as the *exemplar*, which is the most typical and representative actual member within a category. Thus, to discover latent categories, the task can be regarded as discovering the exemplars representing each category.

The work of exemplar detection is closely related to two different areas: *affinity propagation for clustering* and *cluster-based information retrieval*. Under the affinity propagation (AP) algorithm [10], clustering is viewed as the process of identifying a subset of representative exemplars. The AP algorithm is widely deployed in many research fields such as image categorization [37], image segmentation [39] and so on.

A large body of work on cluster-based approach for IR aims for returning a ranked list of a set of exemplar documents representing the clusters of documents relevant to a given query. The research problem here can also be considered as *search result diversification (SRD)*. The *MMR* model [7] applies the *greedy best first strategy* to obtain the ranked list of exemplars. Later on, the Modern Portfolio Theory (MPT) [36] model and the *Exp - 1 - call@k* [30] model were proposed for improving implicit SRD. The well-known Desirable Facility Placement (DFP) [43] model uses the *greedy best k strategy* for ranking the exemplars in a more general way.

Our work is a further endeavor to the exemplar detection paradigm for entity categorization with the difference that we assume a unique history-based approach. Inspired by AP, we propose a concise optimization formulation for selecting exemplars, and we use the *bound-and-branch* method to obtain the optimal solution. The study most related to ours in terms of discovering exemplars is [42], which formulates implicit SRD as a process of ranking k exemplar documents from the top- m documents of an initial retrieval.

2.2 Document Summarization

2.2.1 Timeline Summarization

Timeline summarization defined as the summarization of sequences of documents (typically, news articles about the same events) has been actively studied in the recent years. Yan et al. [40] propose the evolutionary timeline summarization (ETS) to compute evolution timelines consisting of a series of time-stamped summaries. David et al. present

a method for discovering biographical structure based on a probabilistic latent variable model [3]. Tuan et al. [34] introduce a novel approach for timeline summarization of high-impact events, which uses entities instead of sentences for summarizing the events.

2.2.2 Comparative Summarization

Comparative summarization requires the system to provide short summaries from multiple comparative aspects. Wang et al. [35] propose a discriminative sentence selection method based on a multivariate normal generative model aiming to extract sentences best describing the unique characteristics of each document group. Ren et al. [11, 26] explicitly consider contrast, relevance and diversity for summarizing contrastive themes by adopting a hierarchical nonparametric Bayesian model to infer hierarchical relations among topics for enhancing the diversity of themes. In the recent days, differential topic models have also been explored to measure sentence discriminative capability for comparative summarization [14].

2.2.3 Entity Summarization

The main idea of entity summarization is to adapt the presentation of entities toward their individual properties. Thalhhammer et al. [33] present LinkSUM, a new method for entity summarization that follows a relevance-oriented approach to produce generic summaries. Gunaratna et al. [12] explore a novel diversity-aware entity summarization approach that mimics human conceptual clustering techniques to group facts. Entity summarization has also gained attention in industry [4, 31].

Our work is novel in several important aspects: (1) We propose summarization of entire categories of entities rather than summarization of an individual entity, (2) the generated summary is based on the detection of history-based categories and (3) it captures within-category similarities and differences between histories of discovered latent groups, thus enabling history-based category characterization and comparison.

3 Problem Definition

3.1 Input

The input in our task is documents containing descriptions of entity histories. Each such document “spans” over a certain range of time, and each sentence is assumed to refer to a historical event. The dates of events can be either explicitly mentioned in the sentence or they could be estimated based on nearby sentences.

Fig. 2 System overview

3.2 Research Problem

We propose to split the research problem into two subtasks, and we formulate them as follows.

History-based entity categorization Given a set of history-related documents $[d_1, d_2, \dots, d_n]$, each about a particular entity and a preset time window, the task is to detect latent categories $[c_1, c_2, \dots, c_k]$ and their corresponding exemplars $[d_e^1, d_e^2, \dots, d_e^k]$ where entities within each category share similar histories.

Comparative timeline summarization Given a set of history-related documents $[d_1^j, d_2^j, \dots, d_n^j]$, each about a particular entity within the same category j ($j \in [1, \dots, k]$), the task is to select m most important events $[e_1, e_2, \dots, e_m]$ to form a concise timeline reflecting typical history of the category. Each event in the summary is represented by words $[w_1, w_2, \dots, w_i]$. The events to be included in the category timeline should also emphasize the main differences between the history of this category and the histories of other categories.

3.3 Approach Overview

Figure 2 provides an overview of our approach. We first preprocess input documents and extract events. Then, we compute event importance by PU learning. After that, we propose a concise optimization model for the history-based categories' detection. Lastly, we perform a two-layer mutually reinforced random walk for constructing summarized timelines of each category.

4 Event Importance Calculation

Each category is going to be represented by a representative entity (which we call *exemplar*). The characteristics of a category will be thus embodied in the history of its exemplar under the exemplar theory. Naturally, those entities whose histories abound in events important in their respective categories are more representative and they should be chosen as exemplars. Similarly, to generate the *prototype*, which is the constructed summary of the histories of all entities belonging to the same category, we also want to select important events.

The task of estimating historical significance of events is usually done by historians. During this process, several criteria are adopted to help them make judgments: for instance, *remarkable* (the event was remarked on by people at its time or after), *remembered* (the event is important within the collective memory of a group or groups¹), *resulting in change* (the event had significant consequences for the future), and so on. Manually estimating significance of any historical events under the above-listed criteria is of course labor-intensive and time-consuming and may require special expertise; hence, we attempt to estimate the importance automatically. As it is relatively easy to obtain publicly available datasets of historical events marked as important (e.g., the list of important events in 1990s), we propose a method to compute event importance in a semi-supervised way.

Given a set of historical events $P = \{p_1, p_2, \dots, p_k\}$ where each event is marked as important, and a set of unlabeled events $U = \{u_1, u_2, \dots, u_i\}$, the task is to estimate the degree of importance of each event in U . We use $I(u_i)$ to denote

¹ We note that remembering is prone to many biases, yet still event importance is often correlated with the intensity of its remembering within a society.

Table 1 A sample of four historical events with their estimated importance scores from the histories of Japanese cities (dataset D_1 , see Sect. 7.1)

Original sentence	Importance
The tower of Karatsu Castle was built in 1966	0.015483
1949 also saw the opening of Fukushima University	0.362794
January 17, 1995: Great Hanshin earthquake causes more than 100 casualties	0.786335
During World War II, the July 19, 1945 Bombing of Okazaki killed over 200 people and destroyed most of the city center	0.937206

the importance of i th event in U . The key feature of this problem is that there is no negative example set (i.e., labeled unimportant events), which is typically needed for accurate learning of features of important events. In the recent years, PU learning [21] approaches the problem of building classifiers using only positive and unlabeled examples. A few algorithms [20, 41] based on a two-step strategy were proposed for solving the problem as follows.

Step 1: Identifying a set of reliable negative examples from the unlabeled set.

Step 2: Iteratively applying a classification algorithm for generating a set of classifiers and then selecting the best classifier.

In this study, we adopt the 1-DNF technique [41] to identify a set of reliable unimportant events from the unlabeled set U in step 1, and, for step 2, we use the EM algorithm with a NB classifier for constructing the final estimator.

1-DNF 1-DNF technique first constructs a set of positive features containing words that occur more frequently in the labeled important event set P than in the unlabeled set U . Then, an event in U that does not contain any positive feature is regarded as a strongly unimportant event and is included in the reliable negative set RN .

Naive Bayesian classifier Let $C = \{c_1, c_2\}$ be the two predefined classes which represent important events and unimportant events, respectively. Given a set of training events E , we use $x_{e_i, m}$ to denote the word x_i in position m of event e_i , where x_i is a word in the vocabulary $V = \{x_1, \dots, x_{|V|}\}$. The posterior probability $\Pr(c_j|e_i)$ is computed to perform the classification. We have:

$$\Pr(c_j) = \frac{\sum_{i=1}^{|E|} \Pr(c_j|e_i)}{|E|} \quad (1)$$

$$\Pr(x_i|c_j) = \frac{\lambda + \sum_{i=1}^{|E|} N(x_i, e_i) \Pr(c_j|e_i)}{\lambda|V| + \sum_{s=1}^{|V|} \sum_{i=1}^{|E|} N(x_s, e_i) \Pr(c_j|e_i)} \quad (2)$$

where λ is the smoothing factor (typically set as 0.1) and $N(x_i, e_i)$ is the number of times that word x_i occurs in event e_i . Finally, we obtain the NB classifier:

$$\Pr(c_j|e_i) = \frac{\Pr(c_j) \prod_{m=1}^{|e_i|} \Pr(x_{e_i, m}|c_j)}{\sum_{r=1}^{|C|} \Pr(c_r) \prod_{m=1}^{|e_i|} \Pr(x_{e_i, m}|c_r)} \quad (3)$$

EM Each event in P and RN is assigned the initial label 1 and -1 , respectively. Each event $e \in U - RN$ will not be assigned to any label initially but it will get assigned a probability $\Pr(1|e)$ at the end of the first iteration of EM. Then, the set $U - RN$ will participate in EM with its assigned probabilistic labels in subsequent iterations. The EM algorithm consists of the *Expectation* step and the *Maximization* step. In the *Expectation* step, the probabilistic labels of each event $e \in U - RN$ are produced and revised based on Eq. (3), and then the parameters of classifiers are re-estimated in the *Maximization* step using Eqs. (1) and (2). This leads to the next iteration of the algorithm. When EM converges, the degree of importance of each event u in the unlabeled set U takes the value in $[0, 1]$, which is equal to the probability of event to belonging to the class P (i.e., the class of important events). Table 1 shows several examples of events with their estimated importance scores.

5 History-Based Entity Categorization

Having introduced our approach for event importance estimation, we now discuss how entities can be automatically grouped based on similarities of their histories.

We begin by introducing notations used throughout the rest of this paper. For a given set of history-related documents D including n documents $\{d_1, d_2, \dots, d_n\}$ each representing the history of a distinct entity, the task is to detect a partition of D into k latent categories $\{c_1, c_2, \dots, c_k\}$ (k is not known in advance), where the i th category c_i consists of a set of entities which share similar histories and is represented by an exemplar entity $d_e^{c_i}$. In addition, for a certain entity d_i , we use $\{e_1^i, e_2^i, \dots, e_j^i\}$ to represent the events contained in the history of d_i . The similarity between two entities d_i and d_j is then denoted as $s(d_i, d_j)$. The importance and the date of the j th event in the i th entity are denoted by $I(e_j^i)$ and $t(e_j^i)$, respectively.

5.1 Event Representation

We assume a historical event e_j^i to be represented by a sentence and be associated with a date of its occurrence. As not all words of the original sentence are meaningful, each sentence is first normalized by preprocessing steps such as removing stopwords, stemming and retaining the most frequent 5000 unigrams and bigrams. In the recent years,

a more vital role in characterizing histories of entities than trivial events. Given the importance and the date of an event e denoted by $I(e)$ and $T(e)$, respectively, the process is shown in Algorithm 1.

```

input : Two sequences of events  $d_i = \{e_1^i, e_2^i, \dots, e_n^i\}$  and  $d_j = \{e_1^j, e_2^j, \dots, e_m^j\}$ 
output:  $Similarity(d_i, d_j)$ 
1  $M \leftarrow \text{array}[0..n, 0..m]$ ;
2 for  $i \leftarrow 0$  to  $n$  do
3    $M[i, 0] \leftarrow \text{infinity}$ ;
4 end
5 for  $j \leftarrow 1$  to  $m$  do
6    $M[0, j] \leftarrow \text{infinity}$ ;
7 end
8 for  $u \leftarrow 1$  to  $n$  do
9   for  $v \leftarrow 1$  to  $m$  do
10     $cost \leftarrow I(d_i[u]) \cdot I(d_j[v]) \cdot \frac{2}{e^{|T(d_i[u]) - T(d_j[v])|} + 1} \cdot \text{Distance}_{\cosine}(d_i[u], d_j[v])$ ;
11     $M[u, v] \leftarrow \text{cost} + \text{minimum}(M[u-1, v], M[u, v-1], M[u-1, v-1])$ ;
12  end
13 end
14  $Similarity(d_i, d_j) \leftarrow -M[n, m]$ ;

```

Algorithm 1: Weighted Dynamic Time Warping

word2vec [23] was widely utilized for automatically learning the meaning behind words based on neural networks. We use word2vec to represent terms and events. We obtain the distributed vector representations of each word by training the Skip-gram model on the entire English Wikipedia from 2016 using the gensim Python library [25]. The vector representation of an event e_j^i in our case is a weighted combination of the vectors of terms contained in the normalized sentence that describes the event. The weight of a term is its TF-IDF value calculated based on the original corpus of histories of entities within the same category.

5.2 Entity Similarity Calculation

We now introduce the computation of similarity $s(d_i, d_j)$ between entities d_i and d_j where $d_i = \{e_1^i, e_2^i, \dots, e_n^i\}$ and $d_j = \{e_1^j, e_2^j, \dots, e_m^j\}$. Since cosine similarity is not a proper similarity measure for sequences such as sequences of events, we calculate the similarity of two entities by computing the weighted average pairwise similarity of their events. More explicitly, when computing this similarity, events closer to each other and with higher importance score will be more heavily weighted. This is due to the intuition that (a) comparisons between events being far away from each other are less meaningful, and (b) salient events play

5.3 Optimization Model for Exemplar Detection

In this section, we describe the method proposed for categories generation and exemplar detection. Note that finding exemplars means in fact category generation as an exemplar together with entities voting on it constitutes a single category. The goal is to select a subset of entities as exemplars and assign every non-exemplar entity to exactly one exemplar, so as to maximize the overall sum of similarities between entities and their exemplars and the exemplar importance. Then, the objective can be expressed as follows:

$$\text{Obj}_1 = \lambda \cdot \sum_{i,j} s_{ij} h_{ij} + (1 - \lambda) \cdot \sum_i p_i q_i \quad (4)$$

where s_{ij} is the similarity of entity d_i to entity d_j , and p_i denotes the average importance of events contained in the histories of entity d_i . Furthermore, let h_{ij} and q_i be binary hidden variables where $h_{ij} = 1$ indicates entity d_i has chosen entity d_j as its exemplar, and $q_i = 1$ indicates entity d_i is chosen as an exemplar. λ is a trade-off parameter weighting the exemplar importance and the similarities between non-exemplars to exemplars.² Solving Obj_1 is essentially

² We experimentally set the value of λ to be 0.4.

a hard combinational optimization problem, and it can be transformed into a more feasible optimization problem by introducing the following constraint functions:

$$C^1(h_{i:}) = \begin{cases} 0 & \sum_j h_{ij} = 1, \\ -\infty & \text{otherwise} \end{cases} \quad (5)$$

$$C^2(h_{:j}) = \begin{cases} 0 & q_j = h_{jj} = \max_i h_{ij}, \\ -\infty & \text{otherwise} \end{cases} \quad (6)$$

where $h_{i:} = h_{i1}, \dots, h_{in}$ and $h_{:j} = h_{1j}, \dots, h_{nj}$. The constraint function (5) forces each entity to be assigned to exactly one exemplar (which can be itself), while (6) enforces that an entity must be an exemplar if other entities choose it as an exemplar. Thus, the goal is to maximize

$$\text{Obj}_2 = \lambda \cdot \sum_{ij} s_{ij} h_{ij} + (1 - \lambda) \cdot \sum_i p_i q_i + \sum_i (C^1(h_{i:}) + C^2(h_{:i})) \quad (7)$$

By running the *max-sum algorithm* [17], an approximate solution for the above problem can be efficiently achieved. More explicitly, after normalizing all similarities and importance to $[-1, 0]$ range, two sets of messages are calculated iteratively until convergence:

$$\alpha_{ij} = \begin{cases} \frac{p_j(1-\lambda)}{\lambda} + \sum_{k \neq j} \max(0, \rho_{kj}) & i = j, \\ \min[0, \frac{p_j(1-\lambda)}{\lambda} + \rho_{jj} + \sum_{k \notin \{i,j\}} \max(0, \rho_{kj})] & i \neq j \end{cases} \quad (8)$$

$$\rho_{ij} = s_{ij} - \max_{k \neq j} (\alpha_{ik} + s_{ik}) \quad (9)$$

where $\alpha_{ij} = 0$ and $\rho_{ij} = 0$ initially. Intuitively, message α_{ij} corresponds to how willing entity d_j is to serve as the exemplar for entity d_i , while message ρ_{ij} conveys to which extent entity d_i wants entity d_j to be its exemplar. Finally, the exemplar d_j for entity d_i can be obtained by

$$d_j = \arg \max_j \{\alpha_{ij} + \rho_{ij}\} \quad (10)$$

6 Comparative Timeline Generation

In the previous section, we explained the process of dividing entities into latent categories from history viewpoint, where each category is represented by its *exemplar*. The obvious weakness of representing the whole category by a single member entity (i.e., the exemplar) is the risk of missing some key information from other members. The solution to represent a category would be then constructing an “artificial” entity (i.e., a *prototype*) based on all the entities belonging to the same category. Another problem is that

the comprehension of a category’s history should not be done without comparison to the histories of its other categories. To compensate for these two issues, we generate a summarized timeline (effectively, a *prototype*) for each category including not only events similar between the category members but also considering differences from the timelines of other categories. This allows presenting historical knowledge from a comparative aspect and offers effective characterization of events specific to each given category.

6.1 Mutually Reinforced Random Walk

Our task of timeline creation shares some similarities with two special summarization tasks with specific settings: *evolutionary timeline summarization* and *comparative summarization*. On the one hand, the history-related documents we handle have a strong temporal character, and the summary to be created is the evolution trajectory along the timeline consisting of important and time-stamped events, which is consistent with the goal of *evolutionary timeline summarization*. On the other hand, events to be selected into the summary should be *typical* in the histories of entities within the same category and *atypical* in the histories of entities from other categories to reveal category’s unique characteristics.

Let the category to be summarized into a timeline $T(C)$ be denoted by $C = \{d_1, d_2, \dots, d_l\}$. We use $E(C) = \{e_1^E, e_2^E, \dots, e_m^E\}$ to denote the set of all events in the history of all entities within C , and $O(C) = \{e_1^O, e_2^O, \dots, e_n^O\}$ to denote the set of all events belonging to histories of the entities outside C . An event $e \in E(C)$ is more likely to be selected into $T(C)$ if it is similar to events in $E(C)$ while being dissimilar to events in $O(C)$. Furthermore, it is more likely to be included if it is especially similar to typical events in $E(C)$ (rather than similar to any events in $E(C)$) while being dissimilar to typical events in $O(C)$. Based on the above hypotheses, we propose a two-layer mutually reinforced random walk (MRRW) model to generate $T(C)$. The reason why we adopt a MRRW model can be concluded into two reasons. First, its link-based structure makes it easy to incorporate similarity and dissimilarity between event nodes into the way of recursively calculating node importance based on adjacent nodes. Second, it can handle the importance computation of events in both $E(C)$ and $O(C)$ at the same time and can naturally model their mutual influence during this process.

For each category C , we construct a linked two-layer graph G containing event set $E(C)$ and complementary event set $O(C)$. Let $G = (V_E, V_O, Q_{EE}, Q_{OO}, Q_{EO})$, where $V_E = \{e_i^E \in E(C)\}$, $V_O = \{e_i^O \in O(C)\}$, $Q_{EE} = \{q_{ij} | e_i^E, e_j^E \in E(C)\}$, $Q_{OO} = \{q_{ij} | e_i^O, e_j^O \in O(C)\}$ and $Q_{EO} = \{q_{ij} | e_i^E \in E(C), e_j^O \in O(C)\}$. More concretely, an event e_i^E that is in the history of an entity in C is represented

as a node in *event-layer*, and an event e_i^O which belongs to histories of the entities outside C is represented as a node in *complementary-event-layer*. There are three different types of edges corresponding to different relations: event-to-event, complementary-event-to-complementary-event and event-to-complementary-event, which are denoted by Q_{EE} , Q_{OO} and Q_{EO} , respectively.

We then construct event-to-event, complementary-event-to-complementary-event, event-to-complementary-event and complementary-event-to-event affinity metrics L_{EE} , L_{OO} , L_{EO} and L_{OE} . Let $L_{EE} = [\omega_{e_i^E, e_j^E}]_{|V_E| \times |V_E|}$, where $\omega_{e_i^E, e_j^E}$ is from $Sim_{cosine}(e_i^E, e_j^E)$, respectively. We define L_{OO} , L_{EO} and L_{OE} in a similar way, where

$$\omega_{e_i^E, e_j^O} = 1 - Sim_{cosine}(e_i^E, e_j^O) \quad (11)$$

$\omega_{e_i^O, e_j^E}$ is computed similarly as $\omega_{e_i^E, e_j^O}$. Row normalization is performed for L_{EE} , L_{OO} , L_{EO} and L_{OE} . Note that L_{EO} is usually different from L_{OE}^T because of row normalization. We then perform a two-layer mutually reinforced random walk to propagate the node scores based on internal importance propagation within the same layer and external mutual reinforcement between different layers.

$$S_E^{t+1} = (1 - \alpha) \cdot S_E^0 + \alpha \cdot L_{EE}^T L_{EO} S_O^t \quad (12)$$

$$S_O^{t+1} = (1 - \alpha) \cdot S_O^0 + \alpha \cdot L_{OO}^T L_{OE} S_E^t \quad (13)$$

Here, S_E^t and S_O^t denote, respectively, the importance scores of event set V_E and ones of complementary event set V_O at the t th iteration, which integrates the initial score and the score including within- and between-layer propagation. The initial importance scores S_E^0 and S_O^0 of the two sets are the ones computed from the event importance (see Sect. 4) after normalization such that the scores sum to 1. In the experiments, we empirically set $\alpha = 0.85$ as 0.85 is a commonly used value of the damping factor [5].

It can be proved that the closed-form solution S_E^* of Eq. (12) is the dominant eigenvector of M [19], where

$$M = (1 - \alpha) S_E^0 e^T + \alpha (1 - \alpha) L_{EE}^T L_{EO} S_O^0 e^T + \alpha^2 L_{EE}^T L_{EO} L_{OO}^T L_{OE} \quad (14)$$

6.2 Post-processing

MRRW produces summaries in which each event is in the form of a sentence. A sentence may, however, contain too specific details which might be true only for the instance from which it was extracted. Thus, we choose to generalize the top-scored sentences to produce the set of descriptive words representing in a general way a given event type.

Table 2 Summary of datasets (the time ranges of datasets D_1, D_2, D_3 are based on absolute time, while those of datasets D_4, D_5, D_6, D_7 are based on relative time)

Dataset	Wikipedia category	# Entities	Time range
D1	Japanese Cities	532	40—2016
D2	Chinese Cities	357	12—2016
D3	UK Cities	68	1—2016
D4	American Scientists	141	0—103
D5	French Scientists	41	0—101
D6	Japanese PMs (pre WW2)	32	0—98
D7	Japanese PMs (post WW2)	30	0—93

More concretely, for each sentence selected by Eq. (12) indicating an event to be included into the summary, we seek l most similar sentences in the corpus and construct a cluster of $l + 1$ sentences. Then, we compute term frequency-inverse document frequency (TF-IDF) on the created clusters to extract the set of meaningful words describing each cluster. (Table 6 will show examples of such extracted words.)

7 Experiments

7.1 Datasets

We test our methods on entities of different types from different time periods and locations. In particular, we perform experiments on seven Wikipedia categories including three city categories and four person categories. The city categories are Japanese cities, Chinese cities and English cities (denoted by D_1, D_2, D_3 , respectively), while for the person categories we use American scientists, French scientists, Japanese Prime Ministers until the end of WW2 (1945) and Japanese Prime Ministers after WW2 (denoted by D_4, D_5, D_6, D_7 , respectively). Note that, in general, our methods are not bound to Wikipedia categories as any list of entities can form an input, provided the historical description of each entity is available. We use Wikipedia categories in this work as a convenient data source.

For preparing the city categories, each city history is extracted from the “History” section in the corresponding Wikipedia article. To capture historical events, we collect all sentences with dates. As further preprocessing, we reduce inflected words to their word stems and retain only the terms that are among the most frequent 5000 unigrams and bigrams, excluding stopwords and numbers. Each historical event is then represented by the bag of unigrams extracted from its sentence along with the corresponding date.

For the person categories, we utilize a dataset of 242,970 biographies publicly released by Bamman et al. [3]. Every biography consists of several life events, each represented by bag of unigrams and a date. Unlike in the city datasets, the date here is a relative number when counting from a person's birth year. The basic statistics of all the datasets are shown in Table 2.

As for the set of labeled important historical events necessary for event importance calculation, we have collected brief descriptions of key events from Wikipedia year pages³ for each year from AD1 to the present. The total number of the captured event descriptions is 39,881.

7.2 Analyzed Methods

We test our proposed optimization formulation (OM) for history-based categorization and MRRW model for comparative timeline summarization. First of all, to compare with OM, the models *AP*, *MMR* and *DFP* introduced in Sect. 2.1 are used as baseline methods. Apart from the aforementioned methods, we additionally set up a popular clustering technique *K-means clustering* as a baseline. All the baseline methods are briefly discussed below:

1. *Affinity propagation (AP)* [10] views the clustering as identifying a subset of representative exemplars. In particular, it assigns each non-exemplar entity to an exemplar entity under the objective of maximizing the sum of similarities between non-exemplar entities and their assigned exemplar entities.
2. *Maximal marginal relevance (MMR)* [7] is a typical instance of implicit *SRD* (search result diversification) techniques. To obtain the optimal list *L* of exemplar entities, it applies greedy strategy that follows a heuristic criterion of making the locally optimal choice at each round.
3. *Desirable facility placement (DFP)* [43] uses *greedy best k strategy* for generating the desired exemplars' list *L* based on a two-step process. It initializes *L* with an arbitrary solution and then iteratively refines *L* by swapping an entity in *L* with another one outside *L*.
4. *K-Means clustering (K-means)* is a popular method used for cluster detection. It partitions all entities into *k* clusters in which each event belongs to the cluster with the nearest mean to it (given *k* as the size of exemplars).

To assess the proposed MRRW model, we make a comparison with four commonly used multi-document summarization techniques: *LexRank*, *LSA*, *KLSUM* and *MEAD*. The groups used in this step are created by our proposed OM

method. To ensure that the summaries generated by each method reflect also the discrepancy between categories (for achieving comparative summarization), the *MMR* method is adopted for all the methods to penalize the events which appear frequently in other categories. We briefly introduce the baselines used for comparative timeline summarization below.

1. *LexRank* [9] method has been widely adopted in multi-document summarization tasks. It constructs a sentence connectivity matrix and computes sentence importance based on the algorithm similar to PageRank.
2. *LSA summarization (LSA)* is a traditional summarization method that performs latent semantic analysis on terms on sentence matrix as proposed in [32]. It selects semantically important sentences which are different from each other.
3. *KLSUM* [13] is a summarization method that greedily adds sentences to a summary so long as they decrease the KL Divergence between the summary and the document collection. This criterion casts summarization as finding a set of summary sentences which closely match the document set unigram distribution.
4. *MEAD* is a popular extractive summarization method aiming to select a subset of units (e.g., words, sentences) of original documents to form a summary [24]. This centroid-based approach scores sentences based on sentence-level and intersentence-level features including cluster centroid and TF-IDF, etc.

7.3 Experiment Settings

We set the parameters as follows:

(1) *Size of summary*: We experimentally set the summary size of the city datasets and the person datasets to be ten events considering the sizes of their corresponding Wikipedia categories.

(2) *Number of latent categories of each Wikipedia category*: In this study, we use Silhouette analysis [29] to choose a proper value for the number of underlying groups of a given input Wikipedia category. The Silhouette analysis is an example of clustering performance evaluation, where a higher Silhouette coefficient score relates to a model with better clusters. We use K-Means to create the groups based on the number of clusters ranging from 2 to 8, then we pick up the number with the highest Silhouette coefficient score. The estimated number of latent groups for each dataset is shown in Table 3.

³ For example, <https://en.wikipedia.org/wiki/1939>.

Table 3 Estimated number of latent categories of datasets

Dataset	D1	D2	D3	D4	D5	D6	D7
Number of categories	3	2	2	3	2	2	2

7.4 Evaluation Criteria

7.4.1 Evaluation Criteria for Created Categories and Exemplars

For an exemplar to be representative, it should be fairly similar to the other entities within the same category, and there should be a significant difference between it and the exemplars representing other categories. Besides, the chosen exemplars are naturally considered to have salient historical events contained in their histories to fully reflect the key characteristics of their underlying categories. For a given Wikipedia category composed of n entities that become partitioned into k groups $[C_1, C_2, \dots, C_k]$, we use d_i^t and d_e^t to denote the i th entity in the t th group and the exemplar in the t th group, respectively. We then evaluate the representativeness of the identified set of exemplars D_e in terms of the following metrics:

Intra-Similarity (IntraSim) which measures how similar an exemplar is to the entities in its category. The **higher** IntraSim, the more effective the adopted algorithm is.

$$\text{IntraSim}(D_e) = \frac{\sum_{t=1}^k \sum_{d_i^t \in C_t, d_i^t \neq d_e^t} \text{Sim}_{\cosine}(d_i^t, d_e^t)}{n - k} \quad (15)$$

Inter-Similarity (InterSim) which describes how similar an exemplar is to the other exemplars. The **lower** InterSim, the better the performance is.

$$\text{InterSim}(D_e) = \frac{\sum_{t=1}^k \sum_{s=1, s \neq t}^k \text{Sim}_{\cosine}(d_e^s, d_e^t)}{k * (k - 1)} \quad (16)$$

Ratio of intra-similarity to inter-similarity (Ratio) which takes into account both **IntraSim** and **InterSim**, thus reflecting the degree of representativeness of an exemplar.

$$\text{Ratio}(D_e) = \frac{\text{IntraSim}(D_e)}{\text{InterSim}(D_e)} \quad (17)$$

Saliency (AveImp) which measures how important the events in the histories of exemplars are.

$$\text{Saliency}(D_e) = \frac{\sum_{t=1}^k \sum_{i=1}^{|d_e^t|} I(e_{d_e^t, i}^t)}{k} \quad (18)$$

where we use $e_{d_e^t, i}^t$ to denote the i th event of entity d_e^t .

The evaluation of the constructed categories and selected exemplars will be discussed in Sect. 7.5.1.

7.4.2 Evaluation Criteria for Summarized Timelines

Manually creating summaries of histories of categories is difficult. We then ask users to evaluate the quality of generated summaries. We conduct evaluation based on three criteria which are crucial for a high-quality summary. Each event in the summary is graded in terms of:

- *Saliency* which measures how sound and important each extracted event is.

Besides, each summary is graded in terms of:

- *Comprehensibility* which measures how easily the output words can be associated with real events.
- *Diversity* which measures how diverse the events in the summary are (both semantically and temporally).

We have five methods here to be tested (one proposed method and four baseline methods). Five annotators (four males, one female) who have significant interest in history were asked to evaluate the generated summaries. Each summary was ensured to be evaluated by three annotators. During the assessment, the annotators were allowed to utilize any external resources including the Wikipedia, Web search engines, books, etc. All of the scores were given in the range from 1 to 5 (1: not at all, 2: rather not, 3: so so, 4:

Table 4 Performance of different models on the city and person datasets

Model	Data	IntraSim	InterSim	Ratio	AveImp
K-Means	Cities	0.888	0.780	1.139	0.697
	Persons	0.535	0.462	1.190	0.722
AP	Cities	0.884	0.770	1.148	0.732
	Persons	0.619	0.476	1.330	0.779
MMR	Cities	0.820	0.563	1.490	0.898
	Persons	0.572	0.344	2.084	0.879
DFP	Cities	0.874	0.809	1.080	0.912
	Persons	0.770	0.738	1.044	0.879
OM	Cities	0.859	0.543	1.620	0.939
	Persons	0.758	0.478	2.152	0.914

The best result of each setting is in bold

Fig. 3 Parameter tuning

rather yes and 5: definitely yes). After the annotation scores were given, we averaged saliency scores per each summary. Lastly, we averaged all the individual scores given by the annotators to obtain the final scores per each summary (Table 5).

The evaluation of the generated summaries will be discussed in Sect. 7.5.2.

7.5 Evaluation Results

7.5.1 Evaluation Results for Created Groups and Exemplars

We first compare the categorization effectiveness of OM and baselines. Later, we investigate the effects of parameters used in OM, including trade-off parameter λ and the number of latent groups k .

(1) *Categorization effectiveness*: Table 4 shows the performance in terms of *InterSim*, *IntraSim*, *Ratio* and *AveImp*.

We can observe that the proposed optimization framework has better performance than all the baseline methods in terms of two main evaluation metrics, *Ratio* and *AveImp*. Besides, we can also notice that two SRD models *MMR* and *DFP* outperform two clustering models *K-Means* and *AP* in terms of *AveImp*.

Now we investigate the possible reasons for the above results. *MMR* relies on the best first strategy, making it simple and computationally efficient. However, at a particular round, the heuristic criterion may incur *error propagation*. *DFP* can alleviate such problem, but it is based on hill climbing algorithm. A potential problem is that hill climbing may converge to a local maximum. In contrast, our optimization model is able to *globally identify the optimal subset of exemplars*. This largely explains why both *MMR* and *DFP* underperform *OM*. Similar observation has been reported in [42] in the task of Web search result diversification when using *OM*. On the other hand, though *AP* shares

many similar characteristics with *OM*, it does not guarantee to find the optimal solution, hence its lower performance. As for *K-Means*, it suffers from strong sensitivity to outliers and noise, which leads to varying performance.

Both *MMR* and *DFP* models take importance of entities into consideration during the process of identifying exemplars, which can explain why they have better performance in terms of *AveImp* than the two clustering methods. The last finding is that cities may have larger homogeneity than persons, as supported by the observation that generally all methods achieve larger scores in terms of *IntraSim* on city datasets than on person datasets.

(2) *Effects of trade-off parameter*: We examine now how the performance of *OM* varies when we change the parameter λ . We set λ in the range $[0, 1]$ with a step of 0.1. The closer λ is to 1, the less effect the importance part has.

From Fig. 3a, we see that *AveImp* degrades when increasing the value of λ . This observation is consistent with our previous analysis in Sect. 5.3. On the other hand, when λ

Table 5 Performance of different models on city and person datasets

Model	Data	Saliency	Comprehensibility	Diversity
LexRank	Cities	3.457	2.714	3.571
	Persons	4.089	2.333	4
LSA	Cities	3.243	2.571	3.714
	Persons	3.822	2.667	3.444
KLSUM	Cities	3.543	2.429	3.143
	Persons	3.522	2.667	3.556
MEAD	Cities	3.429	2.571	2.571
	Persons	3.878	2.556	3.333
MRRW	Cities	3.686	3.143	4.286
	Persons	4.2	2.556	4.556

The best result of each setting is indicted in bold

Table 6 Events in the summary of Japanese cities (see Fig. 1)

	Event	Terms
1	District	Matsubara, village, district, amami, area part, incorporated, city, tannan, prefectures
	Civil unrest	Occurred, end, widely violent, strike protest, opposed matsukawa, incident, demonstration delayed
	Transportation	Sapporo, route, completed, megumino, built meter, main, linking, highway, bypass building
	Natural disasters	People killed, earthquake, suffered, damage wake, light left, tsunami, mikawa, february, dead, typhoon
	Urbanization	Renamed, irino, neighborhood, hall, split respectively, mura, elevated, status new, incorporated
	Militarization	Japanese, industry, navy, nagoya military, imperial center area, works, warehouse, training, support
	Media	Continued, waraji television, spring largescale, included firebombing, expo, broadcasts, bombing, nhk
	Sports	Shizuoka, held, sport, park, national garden university, pacific, international, high, competition
	Clans	Clan, shimazu, province, local, vassal takada samurai ruled, powerful, perished, lord, unified
	Autonomy	Increased, core, autonomy, system prefectural, government city, establishment designation, structure, place
2	Meiji restoration	Abolition, period kuroda dazaifu, uetsu, reppan part meiji, joined, edo, dispossessed, daimyo
	Battles	Navy, japanese, satsuma, royal, refusal punish previous, pay, indemnity, compensation, charles
	Wars	Japanese, navy, imperial, base, air togos, russojapanese role, orient nickname, nelson naval, military
	Festivals	Festival, first, took, snow, place maple, lantern, held, chrysanthemum cherry blossom, castle, hirosaki
	Construction	Warehouse, stone, torn, stonework, reconstructed original form, dutch, date, constructed, builder
	Transportation	Railway, development, increase, via, scale sagami rapid, railroad, rail, connected, led
	Universities	University, taught, matsue lafcadio learn, author, hirosaki, established
	Commerce	Much, fire, consumes, area, replanned maritime, ginza, commerce, canal, accommodate, city
	Natural disasters	Little earthquake, volcano, throughout, spread, relatively outages, numerous, morioka, hit, extensive
	Shoguns	Daimyo tokugawa, shogun shigeharu, sakamoto rule, position, newly, metsuke, income
3	Wars	War, zenkunen, yoriyoshi, takenori, reinforced dewa, defeated, abe, province, minamoto
	Government	City, suggests, reliable, publicly, point notices, legal, issued, governing, council
	Battles	Summer, battle, ground burned, osaka, sakai
	Commerce	Wealthiest, residents, population, people, living enterprise, earned, commercial, almost, japan
	Christianity	Went, outlawed, hiding escape, christianity, capture
	Merge	Isawa, city, village, modern, merger maesawa, koromogawa, established, district, town
	Christianity	Xavier, prosperity, priests, including, francis documented, christian, sengoku, period, visited
	Missionary	Stand, sent, sendai reach, portugal, padre new, missionary, many, jesuit hour, diogo
	Trade	Yamato trade, using, richest, muromachi, mouth location, inland connect, became, foreign, sengoku
	Business and power	Weaken toyotomi, system stronghold, seized, reportedly power, nobunaga, move, merchant, central, business

For each event we show up to top 10 words due to space limit

increases within the range $[0, 0.5]$, so does the *Ratio* approximately. The value of *Ratio* remains relatively stable when λ is larger than 0.5. Overall, tuning λ has effect on the performance. λ needs to be fine-tuned to achieve an optimal performance. In this study, we set λ as 0.5 based on the observations received from Fig. 3a.

(3) Effects of the number of latent categories:

To analyze how the number of categories k affects the quality of identified exemplars, we investigate how *Ratio* and *AveImp* vary per- k . Figure 3b shows the results. Recall that the number of exemplars is equal to the number of categories and k is set in the range $[2, 8]$.

At first glance, we see that the saliency of exemplars is relatively stable as k varies, as reflected by *AveImp*. Besides, the higher the number of exemplars is, the higher the *Ratio*. A closer look at the trend of *Ratio* reveals that our research

problem is not identical to clustering. This is because the trend of *Ratio* is not consistent with the trend of Silhouette coefficient which we used for estimating the value of k before. As k increases, so does the *Ratio* because the input Wikipedia category is partitioned into smaller categories and, thus, each exemplar is more representative in its group. However, the average degree of representativeness of all entities, which depicts the overall performance of clustering, tends to be higher at smaller k , as reflected by the estimated values of k (see Table 3).

7.5.2 Evaluation Results for Summarized Timelines

Table 5 shows the average scores of summaries in three criteria by all the methods generated on all the city and person datasets, respectively. We first note that *MRRW* model outperforms the baselines based on almost all the criteria.

(The only exception is that *MRRW* achieves worse results than *LSA* and *KLSUM* in terms of comprehensibility by 4% on person datasets.) On average, *MRRW* outperforms all baselines by 16.9% and 11.9% across all the metrics on the city and person datasets, respectively. In particular, *MRRW* achieves better results than all the baselines by 8.1%, 10.0%, 22.7% in terms of saliency, comprehensibility and diversity, respectively.

Now we explore the possible reasons for the above observations. When computing the saliency scores of events in a target group, *MRRW* also utilizes global saliency information from the contrasting group, which may result in a better performance by *MRRW* in terms of saliency. In particular, unlike baselines, in *MRRW*, events similar to the trivial events yet different from salient events of the contrasting groups may also be retained in the summary helping to improve the overall diversity. Finally, *MRRW* intrinsically tends to assign high scores to events specific within one group instead of general events, which may explain its unstable performance in terms of comprehensibility.

7.6 Example Summary

We present in Fig. 1 and Table 6 the summary of three identified latent history-based categories of Japanese cities (dataset D1) using our approach. The summary of each category consists of a timeline containing ten events ordered chronologically (see Fig. 1), followed by a table (see Table 6) which includes up to 15 top-scored words representing each event. For every event in Fig. 1, we display its manually created label based on the extracted terms that are shown in Table 6. In addition, each summary event is associated with two numbers indicating, respectively, the median date and the standard deviation of the occurrence years of the event instances it covers. (Both are computed from the event clusters discussed in Sect. 6.)

As we can notice, cities in category 1 form the largest number of Japanese cities. Most of these cities had their key events quite recently as shown in Fig. 1. In the past, the cities within this group tended to be dominated by powerful local clans, as reflected by the event *Clans*. The modern *transportation* infrastructure in Japan started to advance from the early twentieth century. After WW2 (*Militarization*), these Japanese cities were largely transformed by rapid *Urbanization* and were affected by rapid social development and economic growth, embodied in the events of *Media*, *Autonomy* and *Sports*. During such process, the society once had occasional political protests and violent oppositions, as reflected by *Civil Unrest*. Besides, it can be observed that these Japanese cities often suffered from *Natural Disasters* such as earthquakes, typhoons and tsunamis. *Matsubara* and *Tokyo* serve as good examples of the group.

Histories of cities in category 2 express more features of ethnic culture of Japan. These traditional cities have typical regional characters and local culture, and generally, they are not as modern as the cities in the first category. For example, *Shoguns* was the type of military governor and dominator of Japan for around 700 years until *Meiji Restoration*, which began Japan's transformation from a feudal society to a modern industrialized state, and is regarded as the most significant turning point in the history of Japan. Many cities in this group were involved in continuous *Battles* in the middle and late nineteenth century. Hundreds of Japanese castles were constructed throughout the whole country, as reflected by the event *Construction*. *Festivals* relate to major traditional cultural activities in the lives of city residents, such as viewing the cherry blossom and autumn colors. Cities such as *Dazaifu* and *Kyoto* represent this category well.

Histories of cities in the third category embody the multiethnic side of the Japanese society, by showing the pattern of assimilating foreign culture. As evidence, the corresponding summary includes events such as *Christianity*, *Missionary* and *Trade*.⁴ Cities in this category played a vital role in enhancing the international exchange of Japan in the past, such as *Oshu* and *Nagasaki*.

Finally, some events appear in more than one category, e.g., *Natural Disasters*, *Transportation* and *Wars*. Such events can be regarded as a common characteristic shared by all the categories.

8 Discussions

We discuss here several aspects relevant to the task of history-driven entity categorization and the characteristics as well as the limitations of our approach.

- When calculating the similarity of the histories of two entities by dynamic time warping, we assign higher weights to events closer to each other than to events separated by longer time gaps. This is based on the intuition that correspondences between events being far away from each other are less meaningful and such events should not fall into same cluster. As a result, entities are deemed to have similar history if (a) the events in their histories are semantically similar, and (b) these events are close in the timeline. Furthermore, in the future, we will try to utilize the information of absolute time-stamps of events.
- In this work, the event similarity is based on the cosine similarity between event vectors, where the event vector is the TF-IDF weighted combination of the vectors

⁴ Note that the standard deviations of event occurrence times are 0 here as the total number of used events is quite small.

of terms contained in the event. Here, we assume that word ordering does not affect event similarity, following some previous models on sentence similarity computation such as WMD (Word Mover's Distance) [18] and SIF (Smooth Inverse Frequency) [1]. However, we note that word ordering can be an important factor when computing historical event similarity.

- To prepare historical events for a given entity, we capture all sentences containing dates in the History section of the Wikipedia article corresponding to the entity. We detect the temporal expressions by using spaCy⁵ tool. We adopt this extraction method motivated by the previous work [3, 8]. However, when preprocessing datasets, more refined methods for associating time with sentences could be applied (e.g., [16]).
- We would like to emphasize that the proposed task is a novel kind of historical knowledge generation and organization. Our work makes the first endeavor to propose a optimization formulation for the history-based exemplar detection task. This could offer interesting insights to historians, especially, as professionals could provide more complete data as an input. Furthermore, based on the history-based entity grouping we propose, a history of any given entity could be now seen not independently but rather in relation to the typical history of an underlying latent group it belongs to.
- The method can be extended such that latent groups can be detected for different time periods (e.g., histories of cities during the Renaissance or histories of famous persons during their early careers). Different input time periods will usually result in different discovered latent groups.
- As a prototype is composed of sentences extracted from diverse entities, naturally, coherence of a generated history can be an issue. Currently, we abstract from the extracted sentences by constructing timeline events through representative terms. In the future, abstractive summarization methods could be used.
- A related issue is that prototype events that may contradict each other, though no such cases were observed in the experiments.
- Currently, the exemplars are selected on the basis of the similarities of their histories to histories of other entities. However, other attributes could be also considered in the process of exemplar selection—for instance, popularity or familiarity among users (e.g., while *Dazaifu* may be a good exemplar for its latent group, *Kyoto* which belongs to the same group is more known and recognized by potential users). Hence, entity popularity or importance

could serve as an additional component for the exemplar selection (i.e., used as an additional constraint in OM).

9 Conclusions

It is natural for humans to categorize entities based on their common traits. Given the importance of history on shaping the characteristics of many entities, a useful way to form categories is by considering similarities in entity historical developments.

In this paper, we describe a problem of categorizing entities into history-based categories and constructing their comparative timelines for category characterization and understanding. To solve this problem, we propose an unsupervised approach based on a concise optimization model and multiple random walks. The output summary is in the form of key events represented by the sets of meaningful words along with typical exemplar members. The effectiveness of our methods is demonstrated in experiments on seven Wikipedia category datasets through both qualitative and quantitative analyses.

In the future, we plan to design more problem-specific optimization framework with better scalability based on its intrinsic flexibility for the purpose of entity summarization and understanding.

Acknowledgements This research has been supported by JSPS KAKENHI Grants (#17H01828, #18K19841, #18H03243).

Open Access This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made.

References

1. Arora S, Liang Y, Ma T (2016) A simple but tough-to-beat baseline for sentence embeddings
2. Bairi RB, Carman M, Ramakrishnan G (2015) On the evolution of Wikipedia: dynamics of categories and articles. In: AAAI
3. Bamman D, Smith NA (2014) Unsupervised discovery of biological structure from text. *TACL* 2:363–376
4. Blanco R, Cambazoglu BB, Mika P, Torzec N (2013) Entity recommendations in web search. In: ISWC. Springer, pp 33–48
5. Brin S, Page L (2012) Reprint of: the anatomy of a large-scale hypertextual web search engine. *Comput Netw* 56(18):3825–3833
6. Brooks LR (1978) Nonanalytic concept formation and memory for instances. In Rosch E, Lloyd B (eds) *Cognition and categorization*. Lawrence Elbaum Associates, pp 3–170
7. Carbonell J, Goldstein J (1998) The use of MMR, diversity-based reranking for reordering documents and producing summaries. In: SIGIR. ACM, pp 335–336
8. Duan Y, Jatowt A, Tanaka K (2017) Discovering typical histories of entities by multi-timeline summarization. In: *Proceedings of*

⁵ <https://github.com/explosion/spaCy>.

- the 28th ACM conference on hypertext and social media. ACM, pp 105–114
9. Erkan G, Radev DR (2004) Lexrank: graph-based lexical centrality as salience in text summarization. *J Artif Intell Res* 22:457–479
 10. Frey BJ, Dueck D (2007) Clustering by passing messages between data points. *Science* 315(5814):972–976
 11. Gillenwater J, Kulesza A, Taskar B (2012) Discovering diverse and salient threads in document collections. In: EMNLP. Association for Computational Linguistics, pp 710–720
 12. Gunaratna K, Thirunarayan K, Sheth AP (2015) Faces: diversity-aware entity summarization using incremental hierarchical conceptual clustering. In: AAAI, pp 116–122
 13. Haghighi A, Vanderwende L (2009) Exploring content models for multi-document summarization. In: NAACL. Association for Computational Linguistics, pp 362–370
 14. He L, Li W, Zhuge H (2016) Exploring differential topic models for comparative summarization of scientific papers. In: COLING, pp 1028–1038
 15. Hintzman DL, Ludlam G (1980) Differential forgetting of prototypes and old instances: simulation by an exemplar-based classification model. *Mem Cognit* 8(4):378–382
 16. Jatowt A, Au Yeung CM, Tanaka K (2013) Estimating document focus time. In: Proceedings of the 22nd ACM international conference on information and knowledge management, CIKM '13. ACM, New York, pp 2273–2278. <https://doi.org/10.1145/2505515.2505655>
 17. Kschischang FR, Frey BJ, Loeliger HA et al (2001) Factor graphs and the sum-product algorithm. *IEEE Trans Inf Theory* 47(2):498–519
 18. Kusner M, Sun Y, Kolkin N, Weinberger K (2015) From word embeddings to document distances. In: International conference on machine learning, pp 957–966
 19. Langville AN, Meyer CD (2005) A survey of eigenvector methods for web information retrieval. *SIAM Rev* 47(1):135–161
 20. Liu B, Lee WS, Yu PS, Li X (2002) Partially supervised classification of text documents. *ICML* 2:387–394
 21. Liu B, Dai Y, Li X, Lee WS, Yu PS (2003) Building text classifiers using positive and unlabeled examples. In: ICDM. IEEE, pp 179–186
 22. Mack ML, Preston AR, Love BC (2013) Decoding the brains algorithm for categorization from its neural implementation. *Curr Biol* 23(20):2023–2027
 23. Mikolov T, Chen K, Corrado G, Dean J (2013) Efficient estimation of word representations in vector space. arXiv preprint [arXiv:1301.3781](https://arxiv.org/abs/1301.3781)
 24. Radev DR, Jing H, Styś M, Tam D (2004) Centroid-based summarization of multiple documents. *Inf Process Manag* 40(6):919–938
 25. Řehůřek R, Sojka P (2010) Software framework for topic modeling with large corpora. In: Proceedings of the LREC 2010 workshop on new challenges for NLP frameworks, ELRA, Valletta, Malta, pp 45–50. <http://is.muni.cz/publication/884893/en>
 26. Ren Z, de Rijke M (2015) Summarizing contrastive themes via hierarchical non-parametric processes. In: SIGIR. ACM, pp 93–102
 27. Rosch E (1975) Cognitive representations of semantic categories. *J Exp Psychol Gen* 104(3):192
 28. Roth D, Yih Wt (2005) Integer linear programming inference for conditional random fields. In: ICML. ACM, pp 736–743
 29. Rousseeuw PJ (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J Comput Appl Math* 20:53–65
 30. Sanner S, Guo S, Graepel T, Kharazmi S, Karimi S (2011) Diverse retrieval via greedy optimization of expected 1-call@k in a latent subtopic relevance model. In: CIKM. ACM, pp 1977–1980
 31. Singhal A (2012) Introducing the knowledge graph: things, not strings. Official google blog
 32. Steinberger J, Jezek K (2004) Using latent semantic analysis in text summarization and summary evaluation. In: ISIM, pp 93–100
 33. Thälhammer A, Laserra N, Rettinger A (2016) Linksum: using link analysis to summarize entity data. In: ICWE. Springer, pp 244–261
 34. Tran TA, Niedere C, Kanhabua N, Gadiraju U, Anand A (2015) Balancing novelty and salience: adaptive learning to rank entities for timeline summarization of high-impact events. In: CIKM. ACM, pp 1201–1210
 35. Wang D, Zhu S, Li T, Gong Y (2012) Comparative document summarization via discriminative sentence selection. *TKDD* 6(3):12
 36. Wang J, Zhu J (2009) Portfolio theory of information retrieval. In: SIGIR. ACM, pp 115–122
 37. Wang Y, Chen L (2016) K-meap: multiple exemplars affinity propagation with specified k clusters. *IEEE Trans Neural Netw Learn Syst* 27(12):2670–2682
 38. Woodsend K, Lapata M (2012) Multiple aspect summarization using integer linear programming. In: EMNLP. Association for Computational Linguistics, pp 233–243
 39. Xiao J, Wang J, Tan P, Quan L (2007) Joint affinity propagation for multiple view segmentation. In: ICCV. IEEE, pp 1–7
 40. Yan R, Wan X, Otterbacher J, Kong L, Li X, Zhang Y (2011) Evolutionary timeline summarization: a balanced optimization framework via iterative substitution. In: SIGIR. ACM, pp 745–754
 41. Yu H, Han J, Chang KCC (2002) Pebl: positive example based learning for web page classification using SVM. In: SIGKDD. ACM, pp 239–248
 42. Yu HT, Jatowt A, Blanco R, Joho H, Jose J, Chen L, Yuan F (2017) A concise integer linear programming formulation for implicit search result diversification. In: WSDM. ACM, pp 191–200
 43. Zucco G, Azzopardi L, Zhang D, Wang J (2012) Top-k retrieval using facility location analysis. *ECIR* 7224:305–316