System for Category-driven Retrieval of Historical Events

Yasunobu Sumikawa[*]
University Education Center,
Tokyo Metropolitan University
ysumikawa@acm.org

Adam Jatowt
Dept. of Social Informatics,
Kyoto University
adam@dl.kuis.kyoto-u.ac.jp

ABSTRACT

In this paper, we demonstrate an online system for historical event retrieval. Our system outputs ranked events according to an input text query, time range and category relevance. It is useful for users searching not just for important past events related to input entities but events that belong to specified subset of general categories. It can be also helpful for creating datasets of events falling into specific categories or for generating specialized timelines.

KEYWORDS

Digital history, historical texts, event ranking

ACM Reference Format:
Yasunobu Sumikawa and Adam Jatowt. 2018. System for Category-driven Retrieval of Historical Events. In *JCDL '18: The 18th ACM/IEEE Joint Conference on Digital Libraries, June 3–7, 2018, Fort Worth, TX, USA.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3197026.3203888

1 INTRODUCTION

Studying and analyzing history-related data can provide numerous benefits including improved comprehension of the past and support for finding meaningful connections or analogies over time. One of the common goals of teaching and spreading the knowledge of history is to allow studying how people in the past tried to solve issues and problems, and then apply the acquired knowledge for proposing creative solutions to present issues [4].

Due to the recent information explosion the amount of available data about historical events have been also increasing. Such data includes both the digitized archival documents such as news articles as well as collections of retrospective descriptions of past events (e.g., past events' lists in Wikipedia). This growth demands effective retrieval approaches to let users quickly access what they want. For example, a user (e.g., a journalist or a university student) may be interested in *health and environment* past events involving Japan that took place in the 19th century. He or she may wish to create a specialized timeline representing this particular domain (i.e., health and environment) or may just want to acquire dataset of relevant events for detailed study or as a initial data seed for more extensive retrieval. Yet, such domain-specific event collections are not immediately available on the Web [1].

We propose an interactive search system for searching for historical events by category-based filtering realized through automatic event classification. Our system takes four kinds of data: text query, event category, time range and ranking method. It outputs only event descriptions that include the query terms and those that occurred within the specified time range. More importantly, the returned events are sorted based on their relevance to the selected event category or a set of categories.

Contribution: Compared to related works, the core contribution of our system is to use effective filtering for collecting category-specific range of events.

Singh *et al.* [5] proposed method for supporting historians in searching within document archives. Their approach aims at maximizing coverage and minimizing redundancy with respect to different entities and publication times of documents for a given query. For example, if a person name is given as a query, the method lists news articles reporting the person with different entities and publication times. Sumikawa and Ikejiri [6] proposed approach for temporal analogy retrieval which uses a category-based ranking method. Their search engine performs matrix multiplication to sort past events by the number of categories they share with the categories specified by the user. It assumes however the categories for all the past events are given in advance. In contrast, our search engine provides more search options and is based on automatic event classification using Support Vector Machine (SVM) classifier, thus, it can handle any collections of unlabeled past events.

2 EVENT CLASSES AND EVENT COLLECTION

Event Classes. We adopted nine broad event classes introduced and described in [3] which are based on definitions and guidelines used by Wikipedia editors. The classes are as follows: `Armed Conflicts & Attacks`, `Arts & Culture`, `Business & Economy`, `Disasters & Accidents`, `Health & Environment`, `Law & Crime`, `Politics & Elections`, `Science & Technology` and `Sport`.

Dataset. As an underlying event dataset we collected 70,987 past events from year articles in Wikipedia[2] and Wikipedia's Current Portal[3]. The timespan of the collected events ranges from AD 1 to AD 2016. On average, for all the classes, the descriptions contain 25 words, though the length can be as short as 10 words.

3 SYSTEM DESCRIPTION

Fig. 1 shows the interface of the proposed system[4]. Besides inputting the query, a user can control three options: **event category**, **time range** and **ranking method**. The input query and the selected time range act as filters preparing a subset of events that occurred in the designated time frame and that contain specified query words (typically, name of entities). When no filtering is specified (null

[*] This research was performed while the first author was at Tokyo University of Science, Japan.

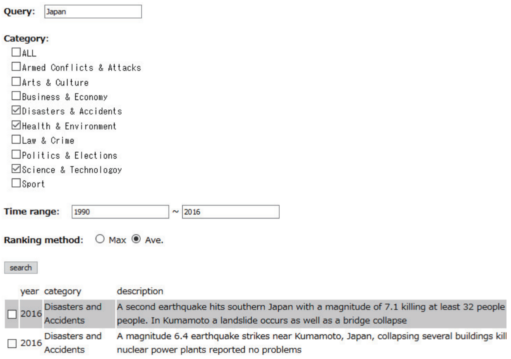
[1] E.g., Wikipedia generally structures important past events chronologically

[2] E.g., https://en.wikipedia.org/wiki/AD_1

[3] https://en.wikipedia.org/wiki/Portal:Current_events

[4] The system can be accessed online: http://tk2-240-29917.vs.sakura.ne.jp/dataset_creation/

JCDL '18, June 3–7, 2018, Fort Worth, TX, USA
© 2018 Copyright held by the owner/author(s).
ACM ISBN 978-1-4503-5178-2/18/06.
https://doi.org/10.1145/3197026.3203888



Query: Japan

Category:

- ☐ALL
- ☐Armed Conflicts & Attacks
- ☐Arts & Culture
- ☐Business & Economy
- ☑Disasters & Accidents
- ☑Health & Environment
- ☐Law & Crime
- ☐Politics & Elections
- ☑Science & Technologoy
- ☐Sport

Time range: 1990 ~ 2016

Ranking method: ○ Max ● Ave.

	year	category	description
☐	2016	Disasters and Accidents	A second earthquake hits southern Japan with a magnitude of 7.1 killing at least 32 people people. In Kumamoto a landslide occurs as well as a bridge collapse
☐	2016	Disasters and Accidents	A magnitude 6.4 earthquake strikes near Kumamoto, Japan, collapsing several buildings kill nuclear power plants reported no problems

Figure 1: Snapshot of system interface.

query and unbounded time) all events stored in the database are considered. The category option allows for ranking the selected events regarding their inherent types. The ranking is based on the degree to which a dedicated classifier (described further in Sec. 3.2) judges an event to be of a particular class. Note that unlike usual approaches in information retrieval (IR) we do not focus on the topical relevance but, instead, we put special emphasis on category relevance. Hence, in the current implementation we use a condition for an event description to represent a relevant event based on simple query containment[5]. Obviously more refined approaches for topical relevance assessment (e.g., semantic ones) can be used instead.

To sum up, event descriptions that contain input query words and that describe events taking place within the specified date range are ranked based on how strongly they are recognized as belonging to the particular set of event categories. The last option, ranking method, determines the way in which the category probabilities are aggregated for ranking event descriptions.

3.1 Classifying Events

To decide event classes, we used SVM with RBF kernel equipped with the following feature groups which are described in detail in [7]: (1) TF-IDF term vectors, (2) LSA vectors (300 dimensions), (3) Doc2Vec vectors (300 dimensions), (4) VerbNet classes of verbs and ones of head verbs, (5) general entity types detected by Yodie [2] and the type of a head entity, (6) TF-IDF weighted vectors of terms captured from articles and their categories as returned by Explicit Semantic Analysis (ESA) [1] which maps target texts to Wikipedia articles. Using feature selection ($k = 2,000$) and 10-fold cross validation our classifier achieved 79.7% of F-measure (Precision of 78.4% and Recall of 79.6%). As labeled data for training we used the Wikipedia's Current Portal collecting 32,362 event descriptions.

3.2 Ranking Algorithm

Our system loads subsets of feature vectors created in Sec. 3.1 only if the occurrence time of their corresponding events matches the time frame T and if the event descriptions contain the input query Q given by the user. The system then applies SVM to the feature vectors in order to estimate to what degree each considered event is relevant to the concerned event classes. The probabilities of classes as estimated by SVM are used as confidence representation[6].

As multiple categories can be input by the user, our system aggregates membership probabilities over different categories. We provide two kinds of aggregation methods: *Max* and *Ave*. The first is the straightforward approach; it ranks events by their maximum confidence among all the selected categories. On the other hand, the second one aggregates based on computing the average confidence. The motivation behind the latter approach is that users may want to collect events belonging to more than one class. For example, the outbreak of Zika virus in 2016 caused death of many people (`Health & Environment` event) but also resulted in the decrease in the population of bees (`Disasters & Accidents` event). To return such events one needs to ensure high confidence probabilities of both the classes. Hence, we provide also an option to calculate average probability values for ranking events.

Finally, our system sorts all the events by their relevant values in descending order and returns them to the user.

More formally, the ranking algorithm is as follows:

$$\begin{aligned} E' &= \{e \mid e \in E \cap \; year(e) \in T \cap Q \in w(e)\} && (1) \\ S(f, e \in E') &= f(CatRel(FV(e))) && (2) \end{aligned}$$

where function $year(e)$ returns a year when the event e occurred, function $w(e)$ returns words used in e and $FV(e)$ outputs a feature vector of e. E is the total set of event descriptions, while E' is the subset of E after the time range and query based filtering. For all the events, our ranking algorithm applies the higher-order function S that takes two arguments, a function f that is either *Max* or *Ave* and the set of feature vectors of e. Then, the algorithm sorts the events by the scores S as being aggregated by f.

4 CONCLUSIONS & FUTURE WORK

In this paper, we demonstrate an interactive online system for retrieving past event descriptions. Our system takes query words, time range and category relevance as an input in order to effectively collect events that related to query, that occurred at a specified time range and, most importantly, that fall into a particular event category. In the future, we plan to propose incorporating more effective methods for representing topical relevance and we plan to add historical importance scores of events.

Acknowledgments. This work was supported in part by MIC SCOPE (#171507010) and MEXT Grant-in-Aids (#17H01828 and #17K12792).

REFERENCES

[1] M.-W. Chang, L. Ratinov, D. Roth, and V. Srikumar. 2008. Importance of Semantic Representation: Dataless Classification. AAAI'08, 830–835.

[2] G. Gorrell, J. Petrak, and K. Bontcheva. 2015. Using @Twitter Conventions to Improve #LOD-Based Named Entity Disambiguation. ESWC'15, 171–186.

[3] A. Košmerlj, E. Belyaeva, G. Leban, M. Grobelnik, and B. Fortuna. 2015. Towards a Complete Event Type Taxonomy. WWW '15 Companion, 899–902.

[4] P. Lee. 2005. Historical Literacy: Theory and Research. *International Journal of Historical Learning, Teaching and Research* 5, 1 (2005), 25–40.

[5] J. Singh, W. Nejdl, and A. Anand. 2016. History by Diversity: Helping Historians Search News Archives. CHIIR '16, 183–192.

[6] Y. Sumikawa and R. Ikejiri. 2015. Mining Historical Social Issues, Vol. 39. IDT'15, 587–597.

[7] Y. Sumikawa and A. Jatowt. 2018. Classifying Short Descriptions of Past Events. ECIR'18, 729–736.

[5] This choice is partly driven by the fact that event descriptions in our dataset are very short. We also assume a user query is going to represent a named entity in most of the cases.

[6] We store the precomputed values offline for online retrieval.