DEIM Forum 2019 P2-123

Present Relatedness Estimation of Archival Documents using External Knowledge Base

Mari SATO†, Adam JATOWT†, and Masatoshi YOSHIKAWA†

† Graduate School of Informatics, Kyoto University Yoshida-honmachi, Sakyo-ku, Kyoto, 606-8501 Japan
E-mail: †msato@db.soc.i.kyoto-u.ac.jp, ††adam@dl.kuis.kyoto-u.ac.jp, †††yoshikawa@i.kyoto-u.ac.jp

Abstract Our society nowadays generates massive amounts of digital data which is being archived and made available for long time. In addition, many born analog documents are being digitized and included in online archives. Although document archives are often used by professionals that often precisely know what they wish to search, when ordinary users search in document archives they often would like to find content which has some relatedness to the present, for example, content related to popular present events or entities. However, search engines for document archives do not consider such aspects and retrieve documents in the same way as search engines in non-temporal collections. In this paper, we propose the notion of "present relatedness" of archival documents - the concept which can be included in retrieval mechanisms for document archives. In particular, we put focus on named entities in documents as their representation and we estimate their relatedness to the present using knowledge bases.

Key words archival search, information retrieval, data mining, text mining

1 Introduction

Our society nowadays generates massive amounts of data which is being archived and made available for long time (e.g. Internet Archive(注1) collecting web snapshots from Alexa). In addition, more and more analog documents are being digitized and included in online document archives (e.g. The Times Digital Archive(注2) and ACL Anthology Reference Corpus(注3)). The continuous development of digital document archives allows users to learn about historical events by searching and browsing old documents as well as reading recent ones about those events. While query suggestion and text indexing methods have been already studied in the context of archival search [1] [2] [3]tran15, relatively little has been done about ranking for archival search.

Professional users such as historians or linguists typically know what they wish to find when searching in or interacting with archives and have good skills in locating relevant documents. On the other hand, general users may have less defined search intents and often would like to find content related to the present. For example, events, figures, or places that are the background for present events. Such content can be not only more relevant and attractive to the users but has also a good chance to be of higher utility. For example, a journalist writing an article on a certain present issue might be more interested in past documents that are related to this issue, rather than ones which have weak connection and relation to the present.

In this paper, we propose the notion of "present relatedness" of archival documents, which indicates their relatedness to the current times. We implement this idea in, particular considering document correspondence to popular present events or entities. We put focus on named entities in documents as their representation and we estimate their relatedness to the present using knowledge bases as well as contents similarity and temporal expressions in the documents. We then train the learning to rank model with documents collected from different periods and then evaluate whether the model can effectively rank documents from a certain time period according to their "present relatedness".

We show two examples of our goal. Figure 1 shows articles from late 1980s. Both tell about economic figures, yet the top article contains a mention of "Donald Trump." Having these two documents, we can say that the top document is more related to the present since Donald Trump is current US president. Figure 2 shows articles from 2000s. Both are related to "France," yet the bottom article tells about protests. Having these two documents, we can say that the bottom document is more related to the present since French protests are more popular topic than French cooking. Our proposed approach aims to rank documents related to the present higher as output.

This paper is organized as follows. In section 2 we discuss

(注1) : https://archive.org/web/
(注2) : https://www.gale.com/intl/c/the-times-digital-archive
(注3) : http://acl-arc.comp.nus.edu.sg/

When it comes to merchandising, Donald J. Trump and Mody Dioum could not be much farther apart. But they agree on one thing: The holiday season was hard on Fifth Avenue street peddlers. And as the avenue stepped back to normal yesterday, it appeared that the city's recent crackdown on merchandise peddlers was still having an effect.

Hale Stores, has moved to coordinate its major businesses by naming Ira Neimark, chairman and chief executive of its Bergdorf Goodman subsidiary, to the additional post of vice president of merchandise development for the Neiman-Marcus Group.

Figure 1 Excerpt news articles from 1987 (top) [17] and 1988 (bottom)(注5).

Jean-Franois Revel, a prolific philosopher, writer and journalist who summoned the classical polemical weapons of Voltaire and Montaigne, including humor, irony and surprise, to illuminate subjects from French cuisine to French anti-Americanism, died on Saturday in Paris.

But Sarkozy, still reeling from massive transit strikes and student protests this month throughout France, is unlikely to use the current unrest as a vehicle to turn introspective or vent his rage too loudly at those he once called "thugs."

Figure 2 Excerpt news articles from 2006 (top) [19] and 2007 (bottom) [20].

related works. In section 3 we define what is present relatedness and we outline our research problem. In section 4 we describe the features and two-stage learning of our method. Then we describe our experiments in section 5 and provide an evaluation of the approach on test data in section 6. We conclude the paper and mention future improvements in section 7.

2 Related Works

A basic search problem in information retrieval (IR) is to estimate standard relevance score of d, given user query q and document d [5]. The purpose of our work is to extend this problem to archival search.

User's information needs for web search have been classified into three groups; informational, navigational, and transactional. Broader analyzed search query logs for the current Web and found out that around a half of the queries are informational in conventional Web search [6]. On the other hand, Costa and Silva analyzed search query logs for an Web archive search engine and found out that navigational need was the most common [7].

One central element of navigational need is temporality of documents. It is reported that a significant number of queries on the Web more or less include temporal intent [10] [11]. By exploiting temporality in archived documents, it can lead to better understanding of old documents and better performance of search engines. There have been research works in the context of archival search recently. Zhang et al. proposed methods to suggest corresponding entities across time / geographical areas as effective queries in archival search [1] [2]. Holzmann proposed an indexing method focusing on anchor texts in a knowledge base [3]. Yet, still little has been done about ranking archived documents with respect to user's search intent. Often, this intent is imprecise, especially, when ordinary users are considered. Yet one can assume that in many cases documents related to present social issues are the ones that the users would wish to find. We then introduce the notion of present relatedness of archival documents and develop initial methods towards its effective estimation. The application of such method can be in the form of a integral component within archival search engines which will let users decide the level of "present relatedness" they wish to have when searching for past documents.

As for the concept of present relatedness, the closest work to ours is probably the one on estimating the importance of historical figures. Jatowt et al. found out decisive features that determine the importance of historical figures using features extracted from link structure, visit logs and article content on Wikipedia [8]. While their work tried to predict the importance score for given historical figures, our work has the objective of estimating how documents are related to the present given query.

3 Problem Statement

Def. 1. *Present relatedness is a relatedness of past documents to the present.*

Then we define our problem as follows: Given user query q and document d published in time T_{past} ($T_{past} \ll T_{now}$), the task is to estimate present relatedness score of d: $PR(d \mid q, T_{past}, T_{now})$. T_{now} is an arbitrary time period. We set T_{now} as [2014, 2018].

4 Proposed Method

4 1 Overview

We first represent documents by a range of features that have high chance to capture their relatedness to the present. We will list the features along with the hypotheses that guide their choice below. Having document representation we then utilize learning to rank models for ranking documents. In particular, we train the model with documents from two different periods whose gap is about 10-15 years and then we

predict and evaluate whether the model can rank documents from the middle of two ages according to their “present relatedness”. In other words, we assume that very old documents are on average less related to recent documents. While it may not always be the case, this assumption holds for a large number of documents. Thanks to this we do not need to manually annotate the data for preparing a training set which would be quite costly. First we explain which features we propose to use to learn the relatedness to the present.

4 2 Entity-related Features

4 2. 1 Importance of each entity in the document

Entities which are just passing mentions in documents are probably of little significance to the task of present relatedness estimation. We then need to select important entities in a document to define later features. We used three simple approaches to find the most important entities in each document.

（1） We counted the frequency of each entity and normalized with the frequency of the whole entities in a document.

（2） We figured important keywords which play the role of describing the document. We computed TextRank score [12] for each entity using the implementation by [13].

（3） We figured the offset of earliest entities appeared in a document.

We then use the five most important entities in each document to compute additional features.

4 2. 2 Activity period of entities

We considered that when a document contains a named entity which is no longer active or no longer strongly remembered, a reader has a lower probability to consider the document to be related to users. Based on this hypothesis, we exploited activity period of each entity. For each entity we want to retrieve the associated years and time intervals. First we collected properties whose data type is date (xsd:date, xsd:dateTime, xsd:gYear etc.(注6)). Then We extracted the DBpedia Linked Data representation with temporal information with properties identifying time intervals (e.g. birthDate/deathDate for figures, and foundingYear/dissolutionYear for organizations). For querying properties and date values, we used the DBpedia SPARQL endpoint(注7). For each document d, we define $NE_d = \{e_1, e_2, \cdots, e_n\}$ as the set of named entities in the document and $APe_i = \{T_{start}, T_{end}\}$ as the activity period extracted from ei. The relatedness of an entity to the present r_{e_i} is defined whether the activity period has overlap with present or not:

$$r_{e_i} = \begin{cases} 0 & (T_{end} < T_{present}) \\ 1 & (otherwise) \end{cases} \quad (1)$$

where $T_{present}$ is an arbitrary number of present year. We set $T_{present} as 2014$.

4 2. 3 Popularity of entities

We considered that when a named entity is not popular enough, a reader has a lower probability to find it relevant and by this find the document related to the present. Based on this hypothesis, we exploited Wikipedia pageviews using Wikimedia REST API(注8) as a convenient measure of popularity of entities. For each entity e_i we retrieved the pageview counts $pageview(e_i)$ of its article. We compute entity $popularity(e_i)$ by normalizing the pageviews with a width of six standard deviations of the pageviews s since it covers the 99.7% of the distribution, eliminating outliers as well (See Equation 2).

$$popularity(e_i) = \frac{pageview(e_i)}{6s} \quad (2)$$

4 2. 4 Connectedness to present knowledge graph

We considered that when a named entity has fewer connections in the present-oriented knowledge graph, it is less related to the present. Based on this hypothesis, we constructed “present” knowledge graph $G(V, E)$, where V is the set of nodes representing DBpedia instances of not only selected entities in the documents but also entities which have internal links with them, and E is the set of edges representing links between V. We constructed G from DBpedia Page Links dataset(注9). We compute the connectedness to the present using the biased random walk theory. Note that $\boldsymbol{R}$ is a vector containing node scores, $\boldsymbol{M}$ is an aperiodic transition matrix, α is a decay factor, and $\boldsymbol{d}$ is the static score distribution vector summing up to one.

$$\boldsymbol{R} = (1-\alpha)\boldsymbol{M} \times \boldsymbol{R} + \alpha \boldsymbol{d} \quad (3)$$

where

$$\boldsymbol{d} = \begin{cases} 0 & (T_{end} < T_{present}) \\ 1/|d| & (otherwise) \end{cases}$$

4 3 Content-related Features

4 3. 1 Cosine Similarity

We measured cosine similarity between each past document and the representative collection of present documents. We used tf-idf, word2vec, and doc2vec as vectorization methods. We explain the details about the collection of present documents in the next section.

4 3. 2 Cosine similarity (Transformed version)

We also measured cosine similarity between each transformed past document and the large collection of present documents. As for the transformation method, we used the one proposed in [1].

(注6) : https://www.w3.org/TR/xmlschema11-2/

(注7) : https://dbpedia.org/sparql

(注8) : https://wikimedia.org/api/rest_v1/

(注9) : https://wiki.dbpedia.org/downloads-2016-10#datasets

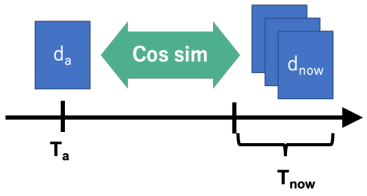

Figure 3 Cosine similarity between present documents.

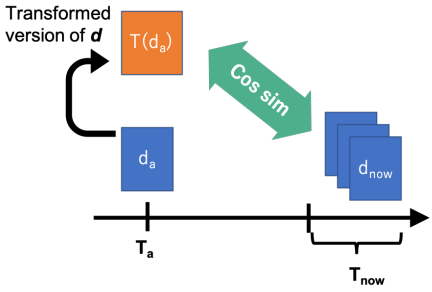

Figure 4 Cosine similarity between present documents and transformed past documents.

4.4 Temporal Expressions

We considered that a when a named entity appeared often together with mentions of past years yet it appears less with mentions of recent years, it is less related to the present. Based on this hypothesis, we extracted years mentions from each document using HeidelTime temporal tagger(注10). We then figured oldest, median, and ealiest year and calcurated their difference between the year $T_{present}$ [14].

4.5 Learning to Rank

Using features described in the previous subsections, we trained the model to learn the present relatedness score of documents. We first gave low score to older documents and high score to newer documents based on assumption that newer documents are more related to the present. We use pairwise approach to train the model to tell which document is more related to the present (See Figure 4). We also use listwise approach to rank documents in order of publication year within collections.

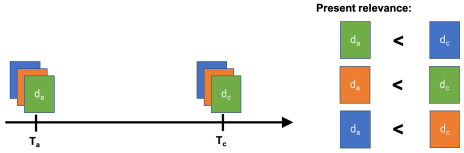

Figure 5 Learning to rank for training.

Using the trained model, we rank documents whose publication dates are not overlapped with documents used in the training process. We randomly selected the ranked lists for evaluation (See Figure 5).

(注10) : https://github.com/HeidelTime/heideltime

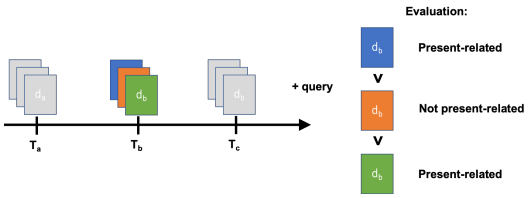

Figure 6 Learning to rank for prediction.

5 Experiments

We sampled 50k articles from the New York Times Annotated Corpus, which were published from 1987 to 1991 and 2003 to 2007 respectively as the collection of old documents. We used them in the learning to rank process. We selected 10k articles from the same corpus, which were published from 1996 to 1997 as the collection of test documents used in the prediction and evaluation process. To extract entity-related features, we first extracted named entities from each document using TextRazor API(注11). The API returns Wikipedia URL with confidence scores given input text. We kept entities with higher confidence score than 0.2. Among kept entities, we assigned importance scores and computed features for five most important entities. We summed up the feature values for a document.

To compute content-related features, we used crawled articles of New York Times from 2014 to 2018 as the collection of present documents.

As for ranking algorithm, we employed LambdaRank [15] in both processes using LightGBM library(注12). We set nDCG as metrics. We included the whole documents from both old and new collections as a ranked list for a query since we did not assume query this time. Next, using the trained model, we rank documents from the test collection according to their "present relatedness". To make evaluation sufficient, we split the collection into genres of news article. We used 142 genres which have more than ten articles. We treated a genre as one query and ranked documents with the same genre.

6 Evaluation

For evaluation, we randomly selected six genres. For each genre, we got 1000 documents returned by Solr. Using the trained model, we re-ranked and pooled top 5 documents. Four annotators gave 1-4 evaluation whether the document is present-related or not (1 and 2 are regarded as not relevant and 3 and 4 are regarded as relevant). Then we evaluated the annotated result by Precision@1, 3, 5 and MAP(See Table

(注11) : https://www.textrazor.com/

(注12) : https://github.com/Microsoft/LightGBM/tree/master/examples/lambdarank

1). As for baseline, we used randomized ranking and cosine similarity. Pairwise approach performed best in Precision@1, 3, and MAP.

Table 1 Evaluation of annotated result. Random and Cosine methods are baselines. Pairwise and Listwise methods are proposed approaches.

	Random	Cosine	Pairwise	Listwise
Precision@1	0.50	0.50	0.60	0.50
Precision@3	0.39	0.53	0.64	0.53
Precision@5	0.42	0.54	0.52	0.46
MAP	0.61	0.63	0.76	0.64

7 Conclusion and Future Work

In this paper, we proposed a method to estimate "present relatedness" of past documents stored in long-term archives using external knowledge base. We put focus on named entities in the documents as their representation and we estimated their relatedness to the present using knowledge bases as well as contents similarity and temporal expressions. As a result, our proposed approach performed better than baselines.

For our future work, we would like to try to indicate why highly-ranked documents are related to the present and thus by this provide explanation to users to help them better understand the returned results. At the same time, we would like to degrade documents with named entities which are not popular enough to be Wikipedia article. These unfamiliar entities are overlooked in our approach.

References

[1] Y. Zhang, A. Jatowt, and K. Tanaka, "Towards Understanding Word Embeddings: Automatically Explaining Similarity of Terms", Proc. of IEEE International Conference on Big Data, pp. 823-832, 2016.

[2] Y. Zhang, A. Jatowt, and K. Tanaka, "Is Tofu the Cheese of Asia?: Searching for Corresponding Objects across Geographical Areas", Proc. of WWW, 2017.

[3] H. Holzmann, Nejdl, and A. Anand, "Exploring Web Achives through Temporal Anchor Texts", Proc. ACM on Web Science Conference, pp.289-298, 2017.

[4] N. K. Tran, A. Ceroni, N. Kanhabua, C. Niederée, "Back to the Past: Supporting Interpretations of Forgotten Stories by Time-aware Re-contextualization", Proc. of ACM International Conference on Web Search and Data Mining, pp. 339-348, 2015.

[5] C. D. Manning, P. Raghavan, and H. Schtze, "Introduction to Information Retrieval", Cambridge University Press, New York, NY, USA, 2008.

[6] A. Broder, "A Taxonomy of Web Search", Proc. of ACM SIGIR forum, pp. 310, 2002.

[7] M. Costa and M. Silva, "A Understanding the Information Needs of Web Archive Users", Proc. of International Web Archiving Workshop, 2010.

[8] A. Jatowt, D. Kawai and K. Tanaka, "Predicting Importance of Historical Persons using Wikipedia", Proc. of CIKM, pp. 1909-1912, 2016.

[9] C. Morbidoni and A. Cucchiarelli, "A Bag-of-Entities Approach to Document Focus Time Estimation", Proc. of KDWEB, 2017.

[10] R. Jones, and F. Diaz, "Temporal Profiles of Queries", Proc. of ACM Transactions on Information Systems, 2007.

[11] D. Metzler, R. Jones, F. Peng, and R. Zhang, "Improving Search Relevance for Implicitly Temporal Queries", Proc. of SIGIR, pp. 700-701, 2009.

[12] R. Mihalcea, P. Tarau, "Textrank: Bringing Order into Texts", Proc. of EMNLP, pp. 404411, 2004.

[13] F. Barrios, F. López, L. Argerich, R. Wachenchauzer, "Variations of the Similarity Function of TextRank for Automated Summarization", Anales de las 44JAIIO. Jornadas Argentinas de Informtica, Argentine Symposium on Artificial Intelligence, 2015.

[14] B. Fetahu, K. Markert, W. Nejdl, A. Anand, "Finding News Citations for Wikipedia", Proc. of ACM International Conference on Information and Knowledge Management, pp. 337-346, 2016.

[15] CJC. Burges, "From ranknet to lambdarank to lambdamart: An overview," Learning pp. 23-581, 2010.

[16] G. Ke, Q. Meng, T. Finley, T. Wang, W. Chen, W. Ma, T. Y. Liu, "LightGBM: A Highly Efficient Gradient Boosting Decision Tree", Advances in Neural Information Processing Systems, pp. 3146-3154, 2017.

[17] "ANTI-PEDDLER DRIVE PLEASES FIFTH AVE. MERCHANTS", The Nwe York Times, https://www.nytimes.com/1987/01/06/nyregion/anti-peddler-drive-pleases-fifth-ave-merchants.html

[18] "CREDIT MARKETS; Neiman Shifts Key Executives", The New York Times, https://www.nytimes.com/1988/04/12/business/credit-markets-neiman-shifts-key-executives.html

[19] "J.-F. Revel, French Philosopher, Is Dead at 82," The New York Times, https://www.nytimes.com/2006/05/02/world/europe/02revel.html, 2006.

[20] "Paris suburb riots called 'a lot worse' than in 2005," The New York Times, https://www.nytimes.com/2007/11/27/world/europe/27iht-riots.4.8500200.html, 2006.