ATAR: *A*spect-Based *T*emporal *A*nalog *R*etrieval System for Document Archives

Yating Zhang
RIKEN AIP Center/NAIST, Ikoma, Japan
Machine Intelligence Technology Dept., Alibaba Group
yating.zhang@riken.jp

Adam Jatowt
Kyoto University
Kyoto, Japan
adam@dl.kuis.kyoto-u.ac.jp

Sourav S. Bhowmick
Nanyang Technological University
Singapore
assourav@ntu.edu.sg

Yuji Matsumoto
NAIST/Riken AIP
Ikoma, Japan
matsu@is.naist.jp

ABSTRACT

In recent years, we have witnessed a rapid increase of text content stored in digital archives such as newspaper archives or web archives. With the passage of time, it is however difficult to effectively perform search within such collections due to vocabulary and context change. In this paper, we present a system that helps to find analogical terms across temporal text collections by applying non-linear transformation. We implement two approaches for analog retrieval where one of them allows users to also input an aspect term specifying particular perspective of a query. The current prototype system permits temporal analog search across two different time periods based on New York Times Annotated Corpus.

KEYWORDS

Temporal analogs, Across-Time search, Document archives

ACM Reference Format:
Yating Zhang, Adam Jatowt, Sourav S. Bhowmick, and Yuji Matsumoto. 2019. *ATAR*: *A*spect-based *T*emporal *A*nalog *R*etrieval System for Document Archives. In *The Twelfth ACM International Conference on Web Search and Data Mining (WSDM '19), February 11–15, 2019, Melbourne, VIC, Australia.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3289600.3290613

1 INTRODUCTION

Keyword-based retrieval is common nowadays and is based on an implicit assumption that searchers are relatively familiar with domains of their search. However, users often need to search in unfamiliar domains such as when looking for information in long-term document archives. Indeed, our knowledge of the past tends to be limited, and an average person typically knows only about major events and entities from the past. Searching within archival document collections as well as understanding the retrieved content can be then hampered due to our limited knowledge of vocabulary that was used in the past and its meaning.

WSDM '19, February 11–15, 2019, Melbourne, VIC, Australia

ACM ISBN 978-1-4503-5940-5/19/02…$15.00
https://doi.org/10.1145/3289600.3290613

To fill in this knowledge gap, methods for finding similar terms over time have been recently proposed [1,2,9,10,11]. In this paper, we demonstrate a temporal analog retrieval system. A user needs to input a term in his/her knowledge domain (typically, the current decade) and the system presents the user with the ranked lists of temporal analogs in the target domain (some time period in the past). Temporal analogs are defined as terms that represent two concepts or entities, each from a different time period, such that they correspond to each other in the sense of playing similar roles or being regarded in a similar way by people in their respective times. For example, for the query `iPod` the system should return words used in the past (e.g., 1980s) that refer to objects having similar functionalities and similar role as `iPod` has nowadays (e.g., `Walkman`). When the query is a person such as `Vladimir Putin`, the analogs' list should include persons who were the leaders of Russia in the past (e.g., `Boris Yeltsin`). Note that a queried term may contain multiple aspects; hence, the temporal analogs depend on a particular viewpoint. For example, `Walkman` corresponds to `iPod` due to similar function of *music device* while `PC` can be a reasonable analog when regarding `iPod` as a *game player*.

Based on the scenarios discussed above, we demonstrate in this paper an end-to-end system called *ATAR* (*A*spect-based *T*emporal *A*nalog *R*etrieval) which unlike previous approaches allows to also input an aspect (also called viewpoint) to enable more accurate temporal analog retrieval. In the system, we first set two time periods: the source and target time, each associated with the collection of corresponding documents published during these times. *ATAR* then provides two methods for the terminology search across time: *General retrieval* that outputs ranked list of analogs for a given query and *Aspect-based retrieval* which allows also for inputting an aspect term besides a query. An aspect term indicates a particular viewpoint based on which temporal analogs should be retrieved (as in the previous example of query `Walkman` with an aspect term *music device*). Our system is unsupervised without the need of any manually annotated data and is able to work on unstructured/raw text data. This indicates the generality of our methodology and possibility to adapt to other scenarios. Furthermore, to facilitate result understanding the system outputs additional supporting data for each result such as representative context in which the returned terms occur. Thus, *ATAR* actually supports a novel type of document archive exploration.

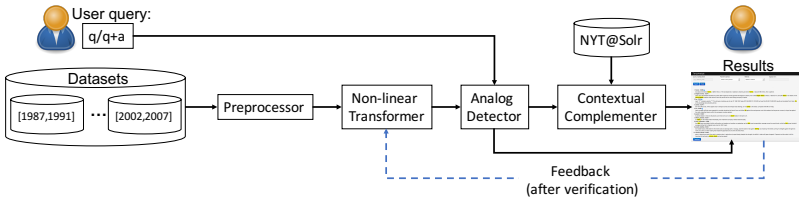

Figure 1: System overview

Related Work. Finding temporal analogs in document archives has recently gained attention of research community. Berberich *et al.* [1] proposed to find semantically similar terms in newspaper archives by using HMM based model and term co-occurrence statistics. Tahmasebi *et al.* [8] detected name changes of input entities by analyzing overlap in their context during change points (e.g., St. Petersburg to Leningrad). Based on that work, the authors have developed a search engine called *fokas* [2] that enriches search results with results for all temporal variants of the query. However, their system is limited to the case of changes in the names of the same named entity. On the other hand, we focus on a broader notion of temporal analogs, hence not necessarily ones used to refer to the same object at different times. More recently, Zhang *et al.* [9, 10] used linear transformation matrix to establish mapping across temporally diverse document collections. Finally, related to the task of temporal analog search is the research on using computational approaches towards diachronic conceptual change (see [7] for a recent overview).

To sum up, our work is novel in several ways. (1) We propose a demo system for end-users to find analogical terms in temporally distant document collections and we contextualize the returned results. (2) We apply non-linear transformation based on neural networks for finding temporal analogs. Finally, (3) we allow users to input an aspect term for biasing the results on a particular perspective and we also propose to collect user feedback.

2 SYSTEM OVERVIEW

ATAR consists of four modules (see Fig. 1 for the overview of system architecture): (1) the *Preprocessor* takes the data (document collections) from the source and target time period. Then for each period, it constructs the distributed feature vector space which contains all the vocabularies existing in the corresponding document collection. (2) The *Non-linear Transformer* is used to conduct a global mapping between the two feature spaces obtained by *Preprocessor*, which can be built off-line. (3) The *Analog Detector* is responsible for taking the input query term (or query and aspect term) and retrieving the ranked list of results with their confidence scores. In the current system, we provide two ranking functions: one built on the global mapping (General Retrieval) and one that considers the bias on a specific aspect (Aspect-based Retrieval). (4) The *Contextual Complementer* provides representative sentences which contain the discovered analogs to offer contextual information for better interpretation of the results.

2.1 Preprocessor

For capturing word semantics we train word embedding vectors separately for each time period obtaining by this two vector spaces (hereafter called source and target spaces). Distributed representation of words by using neural networks was originally proposed in [6]. Mikolov *et al.* [4] improved such representation by introducing Skip-gram model based on a simplified neural network architecture for constructing vector representations of words from unstructured text. We use Skip-gram model as it has advantages of capturing precise semantic word relationships and scaling to millions of words.

2.2 Non-linear Transformer

Our goal is to compare terms in the source space and terms in the target vector space to estimate their similarity and by this to find temporal analogs. Since, we cannot directly compare words in the two different semantic vector spaces, we rely on transformation. Specifically, we train a three-layer neural network (described later) to construct a non-linear transformation for building the basic connection between the vector spaces. Unlike the linear transformation deployed as a transformation matrix [10], the neural network based transformation would allow capturing non-linear mappings across the two vector spaces. For training the network we use a set of training examples called here anchor terms. Note, however that manually preparing sufficiently large sets of anchor terms that would cover various topics/domains as well as exist in any possible combinations of the source and target time periods is infeasible. We then employ here an approximation procedure for automatically selecting anchor pairs. Specifically, we select terms that have high frequency in both the source and the target spaces (e.g., `man`, `city`, `water`). The intuition behind this idea is that terms that are frequent in both the time periods are more likely to have stable meaning and also co-occur with many other terms. The former observation has been validated by linguistic studies of several Indo-European languages including English which demonstrated slower semantic drift of frequently used terms [3, 5]. We note that even if certain anchor term pairs do not retain the same semantics across-time, still, the results should not deteriorate significantly when using sufficiently high number of anchor term pairs.

Suppose there are N pairs of anchor terms $\{(x_1^b, x_1^t),\ldots,(x_N^b, x_N^t)\}$ where x_i^b $(i = 1, \ldots, N)$ (e.g., `man`) is an anchor in source space (e.g., 2010s) and x_i^t is its corresponding anchor, that is, the term with the same literal form (i.e., `man`) in the target space (e.g., 1980s). A three-layer neural network model $\psi(\cdot)$ is established by minimizing the mean square error loss function between $\psi(\mathbf{x_i^b})$ and $\mathbf{x_i^t}$ (see Eq. 1). The activation function at hidden layer is set as hyperbolic tangent ($tanh$), and the dimension of the hidden layer is set to 150. Backpropagation is conducted to perform the training of the neural network with an optimization method of stochastic gradient descent with learning rate of .05.

$$L = \frac{1}{N}\sum_{i=1}^{N}(\psi(\mathbf{x_i^b}) - \mathbf{x_i^t})^2 \quad (1)$$

where,

$$\psi(x) = tanh(\sum_i w_i x) \quad (2)$$

N is the size of anchor term set which contains the top 5%[1] frequent terms in the intersection of vocabularies of the two document sets.

2.3 Analog Detector

General retrieval. After constructing the transformation model $\psi(\cdot)$, we can compute the similarity of a query q in the source space with any term e in the target space. This is done by first predicting the query's vector representation in the target space by $\psi(\mathbf{q})$. Then cosine similarity between the transformed vector and e's vector representation, $\mathbf{e}$, is calculated as:

$$Sim(q, e) = cos(\psi(\mathbf{q}), \mathbf{e}) \quad (3)$$

Aspect-based retrieval. In Aspect-based retrieval, we allow users to give an aspect term and the results are retrieved through biasing candidate temporal analogs based on the given aspect term. This is done by considering not only the semantic but also the relational correspondence.

Formulation. The query tuple τ_{qa} is composed of a base query q and its aspect term a. The objective is to find temporal analog e such that e belongs to the tuple $\tau_{ea'}$ that is most similar to τ_{qa}, where a' is the analog of a.

Tuple similarity computation. To compute similarity between tuples in different vector spaces, we measure both their across-time *semantic* and *relational similarities*.

Semantic similarity is defined as the similarity between the terms in the tuples (similarity of a term in the query tuple to the corresponding term in the candidate analog's tuple). It ensures that the compared terms in the two tuples (i.e., $\mathbf{q}$ compared with $\mathbf{e}$ and $\mathbf{a}$ compared with $\mathbf{a'}$, where $\mathbf{a}$ is vector representation of a and $\mathbf{a'}$ denotes vector representation of its analog) are semantically similar. We use Eq. 4 to compute across-time semantic similarity of tuples:

$$S_{sim}\langle \tau_{qa}, \tau_{ea'} \rangle = \frac{cos(\psi(\mathbf{q}), \mathbf{e}) + cos(\psi(\mathbf{a}), \mathbf{a'})}{2} \quad (4)$$

Relational similarity quantifies the similarity between two relations across-time. Its objective is to measure the similarity degree of the relative positions of terms in relation to their aspect terms. *Relational similarity* is computed as follows.

$$R_{sim}\langle \tau_{qa}, \tau_{ea'} \rangle = cos((\psi(\mathbf{q}) - \psi(\mathbf{a})), (\mathbf{e} - \mathbf{a'})) \quad (5)$$

Lastly, the way to compute the final similarity between the query tuple τ_{qa} and the candidate tuple $\tau_{ea'}$ is:

$$Sim\langle \tau_{qa}, \tau_{ea'} \rangle = \lambda \cdot S_{sim}\langle \tau_{qa}, \tau_{ea'} \rangle + (1 - \lambda) \cdot R_{sim}\langle \tau_{qa}, \tau_{ea'} \rangle \quad (6)$$

In the current implementation, λ equals 0.5. To decrease computational cost for the Aspect-based retrieval, 100 temporal analogs returned by the General retrieval method are used as candidate answers e. a' is set as the most similar analog of a as given by the General retrieval method.

[1] We use 5% as this rate was experimentally verified to result in the best performance for transforming across time periods separated by relatively short time gaps.

2.4 Contextual Complementer

To facilitate the interpretation of the similarity of the retrieved temporal analogs and their meaning, Contextual Complementer module extracts supporting sentence for each returned result from the indexed documents in the target time period. For each analog e, the most representative sentence containing e is retrieved and presented to the user. To discover the representative sentence s, we first construct the set of the sentences which contain e. Then we extract terms that frequently occur within these sentences to form a feature vector weighted by term frequency (centroid vector). Next, we compare each sentence with this vector selecting the one with the highest similarity (the highest cosine similarity score).

3 DEMONSTRATION SYSTEM PROTOTYPE

3.1 Data

For implementing the prototype, we used the New York Times Annotated Corpus containing around 1.8 million newspaper articles published from 1987 to 2007. In the current demonstration, we set the articles published during [2002,2007] as the user's knowledge domain and we provide two past periods to search in: [1992,1996] and [1987,1991]. Each time period contains around half a million news articles, which is sufficient for training word representations. We train the unigram and bi-grams word embedding models of 200 dimensions separately for each time period using Word2Vec. On average, the time periods have 550k unique terms after removing terms with counts smaller than 5.

The document archive has been indexed by Solr and set up on Amazon Web Services (AWS) with Apache Web Server so that documents can be accessed through http requests for collecting candidate sentences.

3.2 Search Interface

Input. The user interface is illustrated in Fig. 2. The query input panel includes four variables: (1) the *Query Term*, (2) the *Past Time Period*, (3) the *Method*, and (4) the *Aspect Term* if the user chooses to use the Aspect-based method. (1) and (4) are passed through text boxes and the rest are implemented by drop-down lists.

Output 1: ranked list of temporal analogs. The main output returned by the model *Analog Detector* is the ranked list of analogs (top 10 terms by default) with their confidence scores computed by either Eq. 3 (General method) or Eq. 6 (Aspect-based method).

Output 2: contextual sentences. As an auxiliary output, the *Contextual Complementer* module extracts the supporting sentence for each returned analog as displayed in Fig. 2. The analog term is highlighted in the sentence for easier recognition.

Feedback Collection. To collect user feedback with evaluations of the retrieved analogs, we added a check box next to each answer which will send back the analog names that are thought to be correct. Then the collected pairs of query and its analog are further utilized as anchor term pairs to train the global mapping function (Eq. 1) after periodically administered manual verification.

3.3 Evaluation

To verify the effectiveness of the proposed methods, we again use the New York Times Annotated Corpus, which has been frequently utilized in the related studies [1, 8, 9]. The query test sets used for testing were taken from [9]. We evaluate the tasks of searching

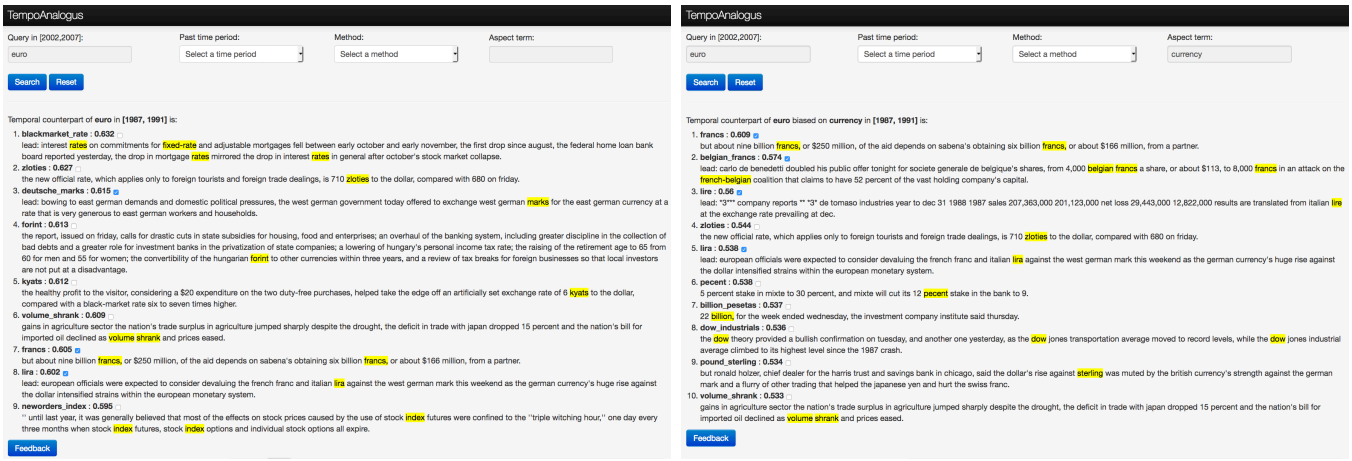

Figure 2: Results for query `euro` in [1987,1991] using General (left) and Aspect-based method (right) biased on term `currency`.

from the most recent period in the dataset [2002,2007] to the two past periods, [1987,1991] (with 95 queries) and [1992,1996] (with 50 queries). We use Mean Reciprocal Rank (MRR) as well as the precision @1, @5 and @10 for evaluating the search results.

Table 1 demonstrates the comparison between the proposed non-linear transformation and the linear approach utilized in the previous works [9, 10]. We can observe that non-linear transformation outperforms the linear one across all the metrics. Note that both methods are evaluated under the same hyperparameter settings (e.g., the size of anchor terms, number of epochs for training).

Table 1: Comparison between linear & non-linear transformation over different time periods.

	[1987,1991]			[1992,1996]		
Method	MRR	P@1	P@5	MRR	P@1	P@5
Linear Tran.	0.298	0.168	0.442	0.161	0.085	0.277
Non-Linear Tran.	**0.309**	**0.173**	**0.456**	**0.173**	**0.091**	**0.283**

To verify the effectiveness of the aspect-based retrieval approach, we manually selected 15 queries for [2002,2007] from the test set, which are the most ambiguous in a sense of having multiple concepts behind them (e.g., `euro`, `smartphone`). We then chose one specific aspect term (e.g., `currency`, `communication`) for each query to conduct search in [1987,1991]. According to the results shown in Tab. 2 we can notice that using aspect terms allows for significant improvements over the general approach in all evaluation metrics.

Table 2: Comparison of General & Aspect-based retrieval on subset of ambiguous terms.

Method	MRR	P@1	P@5	P@10
General Retrieval	0.276	0.143	0.381	0.571
Aspect-Based Retrieval	**0.510**	**0.381**	**0.667**	**0.714**

3.4 Example

Fig. 2 (left) shows an example of searching for the temporal analog of query `euro` in [1987,1991] using General retrieval method. According to the results, we can see that before `euro` was officially adopted in 1995 and became the official currency of the European Union, the EU member states used their own currencies, such as `Francs` for France and `Lira` for Italy. Fig. 2 (right) shows the results of the same query but with bias on aspect *currency* applying the Aspect-based retrieval method. The checked results (see blue boxes after each returned analog) are the ones marked as correct. The comparison between the above two results indicates the results become more precise (the correct answers appear at higher positions) when "narrowing down" to the specific aspect of currency.

4 CONCLUSIONS & FUTURE WORK

We have proposed a demo system for finding similar terms over time to support archival search and for educational purposes. We believe that it can be appreciated not only as a tool for finding temporal analogies, but as a novel means for exploration of document archives, and indirectly, our history and heritage. In future, we plan to extend our approach by automatically selecting the best time periods for finding temporal analogs, as well as use other datasets.

Acknowledgments. This research has been partially supported by JSPS KAKENHI Grant Numbers 17H01828, 18K19841, by MIC/SCOPE #171507010, and by Microsoft Research Asia 2018 Collaborative Research Grant.

REFERENCES

[1] K. Berberich, S. J. Bedathur, M. Sozio, and G. Weikum. 2009. Bridging the Terminology Gap in Web Archive Search *(In Proc. of WebDB)*.
[2] Helge Holzmann, Gerhard Gossen, and Nina Tahmasebi. 2012. fokas: Formerly Known As–A Search Engine Incorporating Named Entity Evolution. *Proceedings of COLING 2012: Demonstration Papers* (2012), 215–222.
[3] E. Lieberman, J. B. Michel, J. Jackson, T. Tang, and M. A. Nowak. 2007. Quantifying the evolutionary dynamics of language. *Nature* (2007), 713–716.
[4] T. Mikolov, K. Chen, G. Corrado, and J. Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proc. of ICLR Workshop*.
[5] M. Pargel, Q. D. Atkinson, and A. Meade. 2007. Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature* 449 (2007), 717–720.
[6] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1985. *Learning internal representations by error propagation.* Technical Report. California Univ, San Diego La Jolla Inst. For Cognitive Science.
[7] Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of Computational Approaches to Diachronic Conceptual Change. *arXiv:1811.06278v1* (2018).
[8] N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. 2012. NEER: An Unsupervised Method for Named Entity Evolution Recognition *(In Proc. of Coling)*. 2553–2568.
[9] Y. Zhang, A. Jatowt, S. S. Bhowmick, and K. Tanaka. 2015. Omnia Mutantur, Nihil Interit: Connecting Past with Present by Finding Corresponding Terms across Time. In *Proc. of ACL*. 645–655.
[10] Y. Zhang, A. Jatowt, S. S. Bhowmick, and K. Tanaka. 2016. The Past is Not a Foreign Country: Detecting Semantically Similar Terms across Time. *IEEE Transactions on Knowledge and Data Engineering* 28, 10 (2016), 2793–2807.