DiaQueTT: A Diachronic and Queryable Topic-Tracking Model

Yuta Nakamura
Kyoto University
Yoshida-Honmachi, Sakyo
Kyoto, Japan
y-nakamura@db.soc.i.kyoto-u.ac.jp

Yasuhito Asano
Toyo University
28-20, Shirayama 5 chome, Bunkyo
Tokyo, Japan
yasuhito.asano@iniad.org

Masatoshi Yoshikawa
Kyoto University
Yoshida-Honmachi, Sakyo
Kyoto, Japan
yoshikawa@i.kyoto-u.ac.jp

ACM Reference format:
Yuta Nakamura, Yasuhito Asano, and Masatoshi Yoshikawa. 2019. DiaQueTT: A Diachronic and Queryable Topic-Tracking Model. In *Proceedings of SIGIR 2019 Workshop on ExplainAble Recommendation and Search, Paris, France, July 25, 2019 (EARS'19),* 5 pages.
DOI: 10.1145/nnnnnnn.nnnnnnn

ABSTRACT

Digital archives generally contain large amounts of documents over a long period of time and diachronic polysemies, which are words that have changed their semantics over many years. We propose a new model of diachronic queryable topic detection and tracking, named DiaQueTT, for digital archives considering these words. DiaQueTT creates a canonical semantic space of both words and documents through many years by extending diachronic word embedding methods, and employs von Mises-Fisher distribution for topic detection and tracking which is able to explain the difference of topics. We evaluate DiaQueTT using the AMiner Citation Dataset containing more than 15 million research papers. DiaQueTT outperforms LSI in terms of the query processing and outperforms LDA and DTM in terms of topic detection accuracy and coherence. We also indicate that DiaQueTT can provide information about the direction of topic changes.

1 INTRODUCTION

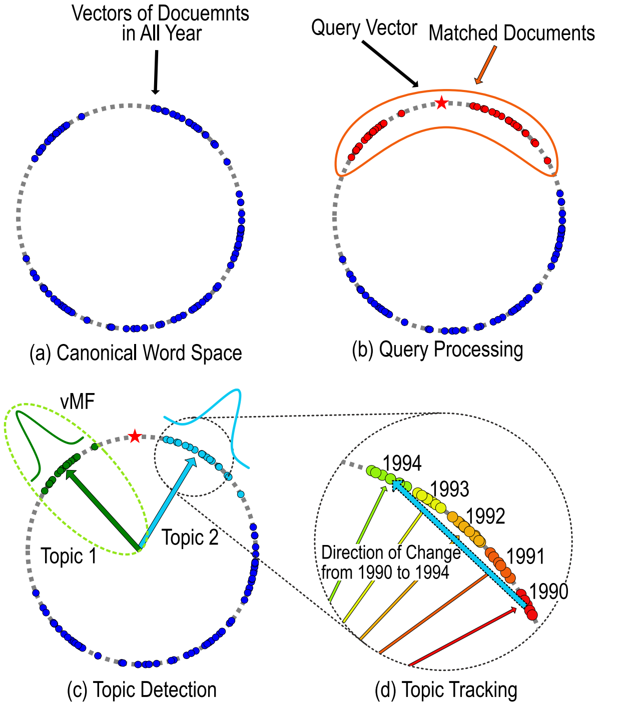
Digital archives have been widely used for various kinds of documents such as academic papers and news articles. These digital archives generally contain large amount of documents over a long period of time; thus, it is difficult for users to survey their interested topics and track the changes in topics over time. In particular, the transition of word semantics over a long period makes such topic detection and tracking even more difficult. For example, the word "cloud" mainly represented clouds in the sky in the 1990s, but the current major semantics of "cloud" is cloud computing; such a word is called "diachronic polysemy." Therefore, a system that satisfies the following conditions would be useful for digital archives: (1) it can manage the transition of word semantics including diachronic polysemies, (2) it accepts a query that consists of a pair of a word and a year, and (3) it achieves topic detection and tracking among the documents that match the query. Let us consider an example that a user wants to survey cloud computing from an academic paper database. The user inputs pair ("cloud", the current year) as a query to the system. Then, the system presents the topics found in the papers that are matched to the query and how they have transitioned. The transition include the following: "there was research on Linux clusters as a past related to cloud computing, and research topics have slightly shifted from efficiency to security in the cloud computing recently." To realize such a system, we propose a new topic detection and tracking model, which is called diachronic and queryable topic tracking (DiaQueTT).

The shift of word sense is actively researched [1–4] but none of them is applied to topic detection. Although some researches [5, 6] deal topic detection over time, they didn't consider the shift of word sense. Some works about the shift of word sense leverage diachronic word embedding [7, 8] which enables us to explain how the word sense shifted over time. To achieve the condition (1), we can utilize diachronic word embedding that assigns a semantic vector to a pair of a word and time. That is, the vectors (Linux cluster, 1995) and (cloud, 2015) are close to each other, but the vector (cloud, 1995) is far from them. However, this technique cannot be directly used for our purpose; for query processing (condition (2)), it is necessary to place words and documents in the same semantic vector space. Therefore, by extending the idea of diachronic word embedding to documents, we propose a novel concept, "canonical semantic space", which enables the transition of word semantics to be managed and integrates words and documents in the same space. For example, by inputting the query (cloud, 2015), we can collect documents about cloud computing including those related to its ancestor concept such as Linux clusters. This is the first contribution of our work.

Among the document set collected for a query as explained above, we perform topic detection and tracking (condition (3)). For example, there are several topics over time about "cloud" such as "performance" and "infrastructure". For this purpose, we propose an approach to represent topics by a mixed von Mises-Fisher (vMF) [9] distribution on our canonical semantic space. This is the second contribution of this study. This approach has three advantages: (a) it enables topic tracking considering the shift of semanti in the meaning of words, including diachronic polysemies such as "cloud", because of the canonical semantic space, (b) it is a natural method of topic detection because a mixed vMF corresponds to a mixed Gaussian distribution of angles on an n-sphere; the mixed Gaussian distribution is widely used for clustering, and (c) it captures the direction of a topic change by observing the trajectory of the "mean vectors" of the distribution over time; for the example above, the mean vector of the distribution that corresponds to articles about the distributed system in 1990 may be close to the query vector (mainframe, 1990), while that in 2015 may be close to the query vector (cloud computing, 2015). Then, the direction of this change can be represented by the difference of these two vectors. This

EARS'19, Paris, France

DOI: 10.1145/nnnnnnn.nnnnnnn



(a) Canonical Word Space

(b) Query Processing

(c) Topic Detection

(d) Topic Tracking

Figure 1: Summary of DiaQueTT in 2 dimensions. (a) A gray broken line means 2-dimensional vMF. The blue points indicate document vectors, and the vectors for all years are located on the same sphere. (b) Documents with high similarity with a given query (*word*, *year*) are selected. (c) From the selected documents, vMF retrieves the topics. (d) For each topic, the changes in each year are tracked, and the direction is calculated.

vector of the direction of the change will be explained by the words which have high similarity with the vector.

Kayhan et al. [10] proposed a topic model with extended LDA (latent Dirichlet allocation) [11] using vMF. Although the performance of the model is good, it cannot achieve (a) and (c). The dynamic topic model (DTM) [12] enables topic tracking by word distributions, while it cannot handle diachronic polysemies. Furthermore, these models cannot handle user queries (see condition (2)).

We evaluate DiaQueTT using Aminer Citation Dataset [13] of academic papers. We perform two case studies and three kinds of quantitative evaluations. One of the three quantitative evaluations is the performance of query processing compared with LSI (latent semantic indexing) [14]. The second evaluation is the accuracy of topic decomposition compared with LDA and DTM, and the third one is the coherence of the detected topics. The experiments also demonstrate that DiaQueTT outperforms competitive approaches in terms of handling diachronic polysemies.

2 DIAQUETT MODEL

In this section, we describe our DiaQueTT model. Fig. 1 (a) illustrates the canonical semantic space construction phase that converts word and documents in all periods to vectors on a n-sphere. We first construct a word space for each year by applying word2vec [15] to a set of documents that have been published in that year. At this point, the word spaces obtained from different years cannot be comparable. Therefore, we construct a transition matrix that specifies which point in a word space corresponds to a given point in another word space using a method proposed by Hamilton et al. [7]. Although it is possible to obtain a comparable word space by continuously learning sentences separated by time, the accuracy is low when the time between two word spaces are far so we leverage the method. [7]. In summary, we obtain a canonical word space for all years: for arbitrary pair (w, y) of word w and year y, e.g., (cloud, 2015), corresponding point $v(w, y)$ in that space is obtained. We also put each document d_i as point $\mathbf{d_i}$ in this space considering its published year y. We define $\mathbf{d_i}$ as a weighted average of vectors $v(w, y)$ of all pairs of words $w \in d_i$ in the document and y, whose weight is $\mathrm{TFIDF}_y(w)$, the TF-IDF of w in the documents published in y:

$$\mathbf{d_i} = \mathrm{normalize}\left(\sum_{w \in d_i} \mathrm{TFIDF}_y(w) \cdot v(w, y)\right),$$

where $\mathrm{normalize}(\alpha) = \alpha/||\alpha||$. We also tried two other candidates to convert a document into a vector. One is Doc2vec [16], and the other uses the simple average of the vectors of contained words. According to our preliminary experiment, the TF-IDF weighted method is the most robust.

Fig. 1 (b) illustrates the query processing phase. When a user want to understand the topic transition about a query, it is required to perform query processing to search for past documents and extract the topics from the query result.

Conventional query processing is not sufficient for our purpose because we need to extract documents that are related to the query even if the query word has not appeared in those year. Therefore, our model treats a pair of the query word and a year as the input, such as (cloud, 2015), and calculates the similarities between each document and the vector that corresponds to the pair in the canonical semantic space. In this way, we can find similar documents to the query, even in the time when the query word has not appeared.

Fig. 1 (c) illustrates the topic detection phase. We assume that the vectors of documents related to a topic are normally distributed around the vector that represent the topic in a word space. We use a vMF distribution to represent a topic because it corresponds to a normal distribution on an n-sphere; it has the mean vector μ and concentration parameter κ, which correspond to the mean and variance in a normal distribution, respectively. Note that a multivariate normal distribution is not suitable because its assumption of normality on each dimension is meaningless in a word space. The probability density distribution $p(\mathbf{d_t} \mid \Theta)$ generated by a certain document d_t from vMFs that represents M topics is:

$$p(\mathbf{d_t} \mid \Theta) = \sum_{m=1}^{M} \pi_m f(\mathbf{d_t} \mid \mu_m, \kappa_m),$$

$$f(\mathbf{d_t} \mid \mu_m, \kappa_m) = C_n(\kappa_m) \exp(\kappa_m \mu_m^T \mathbf{d_t}),$$

$$C_n(\kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)}$$

where Θ is a parameter of a linguistic model and π_m is the mixture ratio for the m-th distribution and $I_{n/2-1}$ the modified Bessel function (see [9]) of the first kind at order $(n/2 - 1)$. We here employ the

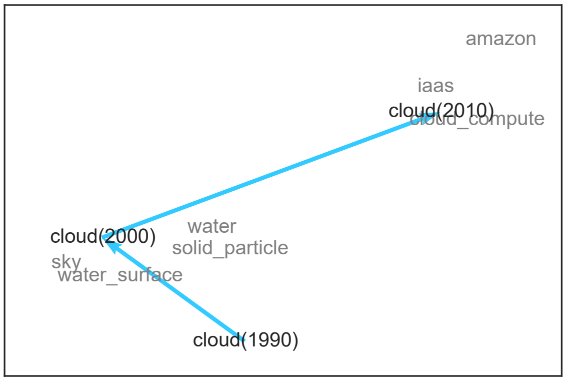


Figure 2: Detection of Changes in "cloud" Semantics

EM algorithm which is commonly used for determining parameters π_m, μ_m, and κ_m. In this phase, first, we represent the document vectors obtained for all periods in the previous phase as a vMF mixture distribution and consider it as the set of topics related to the query. Then, we consider each component (a vMF distribution that corresponds to a subset of documents) of this mixture as a topic. The mean vector of a component is considered the representing vector of its topic. The number M of topics is determined by the Bayes information criteria.

Fig. 1 (d) illustrates the topic tracking phase. To provide how each topic changes over time, we first divide its corresponding subset of documents into small subsets year by year. An example of a divided document sets is represented by the dots with labels 1990, 1991, . . ., 1994. Then, we fit a vMF distribution to the document vectors of each year. The obtained mean vector of each distribution represents the topic in the year. Therefore, the difference between two mean vectors, which is represented by an arrow in the figure, is considered the manner in which the topic changes. This model of topic change has two advantages. First, one can find words that represent the change by searching similar words to the direction of the change, that is, the difference between two mean vectors because word2vec has compositionality of words. Second, one might be able to predict a topic change in the future by analyzing the direction.

3 EXPERIMENTS AND EVALUATION

We experimentally evaluate DiaQueTT using a collection of 1,570,902 papers from 1990 to 2014 in the data set of the citation network provided by AMiner [13]. First, we conduct two case studies: (i) the word semantic changes and (ii) topic tracking for a query. Second, we conduct three quantitative evaluations: (a) the quality of the query result compared to LSI, (b) the quality of topic detection compared to LDA and DTM, and (c) the quality and coherence of the entire topic detection compared to LDA and DTM.

We use the word "cloud" in the case study (i) for verifying that our model can detect the transition of the meaning of words as in previous research. For the word "cloud", it is desired to capture the transition that the term is shifting from the meaning of cloud in the sky to the meaning of cloud computing. t-Stochastic neighbor embedding (t-SNE) [17] is often used when compressing word space such as word2vec into low dimensions for visualization. t-SNE requires a large number of words for better visualization, so we use the most frequent 500 words. Fig. 2 illustrates how the word "cloud" moves in the word of 2014 using t-SNE. The semantics of the word changed from cloud in the sky to the concept of cloud computing, which corresponds to the history.

Table 1: Entire Topics of "machine_learning".

	Similar Words
Topic 1	data_mining, natural_language_processing
Topic 2	classifier, classification

Table 2: Meaning of topic changes about "machine_learning".

	1995 to 2005	2005 to 2014
Topic1	unlabeled_datum, social_medium_website	apis, sentiment_analysis
Topic2	web_log, urls	introductory_material, advanced_undergraduate_student

For the case study (ii), topic tracking for a query, we use (machine_learning, 2014) as a query. As a result, two topics are detected according to our DiaQueTT model using the entire paper collection as a corpus. Table 1 shows similar words to the mean vectors of the two obtained topics. As we explained above using Fig.1 (d), we also obtained the mean vector for each year in each topic. Because the difference between the mean vectors of two different years represents the direction of topic change, showing words similar to the difference vector may help the users to track the topic change. The words similar to the difference vector from 1995 to 2005 and that from 2005 to 2014 are shown in Table 2. Topic 1 on text mining has unlabeled data and social media from 1990 to 2005, and it incorporates the application programming interface (API) and semantic analysis from 2005 to 2014. Topic 2, which is related to classifiers that incorporate concepts on the web from 1995 to 2005 and incorporates regarding the introduction from 2005 to 2014. This result may reflect the fact that the number of papers about machine learning introduction for nonspecialists is increasing with commoditization. Thus, the results appear reasonable. Conventional topic tracking shows the words with high likelihood of each topic for each year to help the users understand the change in the topic. This approach is insufficient because the meaning of a word itself can be changed. In addition, high-likelihood words sometimes have not changed over the years even if a small change in topic has occurred. DiaQueTT can capture such a change by calculating the difference of the mean vectors over years.

Now, we conduct the quantitative evaluation (a) of our DiaQueTT and LSI for query processing to confirm whether each model accurately calculates the similarity between documents because the similarity is a basis of topic modeling. LSI is a good baseline here because it is known to have stable performance in this criteria. We used five queries: "cloud", "hadoop", "android", "dos" and "lsi", which are diachronic polysemies or new words. Note that LSI

Table 3: MAP score of query processing from 1990 to 2014.

	cloud	android	hadoop	dos	lsi
DiaQueTT	**0.8**	**0.72**	**0.72**	**0.85**	**0.94**
LSI	0.60	0.62	0.52	0.09	0.85

Table 4: MAP score of query processing from 1990 to 2000.

	cloud	android	hadoop	dos	lsi
DiaQueTT	**0.77**	**0.55**	**0.63**	**0.74**	**0.91**
LSI	0.43	0.24	0.34	0.00	0.78

Table 5: Performance of topic detection.

	cloud	android	lsi	dos	sns
LDA	0.03	0.01	0.24	0.04	0.01
DTM	0.02	0.07	0.27	0.32	0.14
DiaQueTT	**0.7**	**0.79**	**0.93**	**0.6**	**0.62**

learns the lemmatized data similarly to DiaQueTT, in 300 dimensions. We used sets of papers in two different periods: 1990-2014 and 1990-2000. We retrieved five papers for each year with the highest similarity obtained by each model. For example, we retrieved 125 papers in the data set for 1990 to 2014 (25 years) using each model. The retrieved papers were judged whether the documents are close to the query, by Master or PhD course students who have sufficient knowledge about the topics.

Tables 3 and 4 show the MAP (mean average precision) [18] for the data sets of 1990-2014 and 1990-2000, respectively. The query results of our model show higher accuracy than LSI for both the data sets. In particular, even before the appearance of a query word (in the 1990-2000 data set), we can find papers similar to the query, and our model outperforms LSI in terms of query processing. Now, we quantitatively evaluate (b) DiaQueTT, LDA and DTM for topic detection which are widely used as strong baselines. DiaQueTT is expected to enable the users to detect topics with arbitrary granularity, especially the topics in the result as we explained using Fig. 1 (c). To confirm this expectation, we prepare the ground truth for a pair (w, y) of query word w and year y as follows. First, we randomly select 10 papers that contain w in each year. If there are not sufficient documents, we use all of them. Then, each paper is regarded to belong to the topic represented by w in year y if both of two annotators (students as above) agreed. Specifically, if a given pair is (cloud, 2014), then the papers related to the infrastructure of computers are considered to belong to the desired topic. Consequently, we evaluate how correctly each model gathers the papers that belong to the topic into a cluster. We obtain two clusters by each model for (w, y), one is expected to the cluster of belonged papers and the other is expected to remaining papers. We compare the result with the ground truth using the adjusted Rand index [19]. Since DTM is a time-series expanded model, learning is conducted by separating each year. In DiaQueTT and LDA, we use the the same input data without dividing the data of 25 years for each year.

Table 5 indicates DiaQueTT outperforms the other methods. There are two reasons for this result: (1) Since DiaQueTT classifies the documents of 25 years simultaneously considering semantic changes, it can correctly classify the documents even if the usage of words changes. (2) Since DiaQueTT uses all documents for creating the word vectors, it can classify them without deteriorating the accuracy even if the document set to be classified is small.

Table 6: Coherence of topics using the entire dataset.

Algorithm \Topic number	20	50	100
LDA	0.43	0.41	0.37
DTM	0.40	N/A	N/A
DiaQueTT	**0.61**	**0.61**	**0.60**

Next, we conduct a quantitative experiment (c) to evaluate the performance in common IR tasks. We obtain topics from the entire corpus of 1990-2014 according to each model and evaluate their quality using the topic coherence [20], which is a frequently used measure in such tasks. We use a set of 10 words with high likelihood for each topic for LDA or DTM, and a set of 10 words that are close to the mean vector of each topic for DiaQueTT.

To calculate the coherence, we calculate the normalized pointwise mutual information (PMI) for the entire corpus. The average coherence values of DiaQueTT, LDA and DTM are shown in Table 6. DTM could not return the result for 50 and 100 topics in practical time (it took 40 days for 20 topics). DiaQueTT gives the best score for any number of topics. It is known that word2vec learns features similarly to PMI. DiaQueTT is considered to preserve this advantage of word2vec even when it outputs words that represent a topic in the same manner as conventional topic modeling.

4 CONCLUSION AND DISCUSSION

We propose a new model of queryable topic detection and tracking considering diachronic polysemies for digital archives. Our model solves the problem which the any previous model could not consider the shift of word meaning and outperforms LSI in terms of query processing and LDA and DTM in terms of topic detection and coherence. Our model can be a fundamental technique applicable for many other IR tasks. For example, by adopting "documents" as a query, we can retrieve words and documents related to the query documents even if the meanings of some words have changed. Our model is applicable for labeling such as "cloud computing" to a large set of documents and provides articles with the foresight which have high similarity with a topic in present. Additionally, because our model represents the topics, their changes and trajectories as vectors which could explain their meanings, prediction of the direction of topic changes might be possible as future work.

ACKNOWLEDGEMENT

This work was supported by JSPS KAKENHI Grant Number No. 18H05031

REFERENCES

[1] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten De Rijke. Ad hoc monitoring of vocabulary shifts over time. In Proceedings of the 24th ACM international on conference on information and knowledge management, pages 1191–1200. ACM, 2015.

[2] David Bamman and Gregory Crane. Measuring historical word sense variation. In Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries, pages 1–10. ACM, 2011.

[3] Hosein Azarbonyad, Mostafa Dehghani, Kaspar Beelen, Alexandra Arkut, Maarten Marx, and Jaap Kamps. Words are malleable: Computing semantic shifts

in political and media discourse. In Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, pages 1509–1518. ACM, 2017.

[4] Nina Tahmasebi and Thomas Risse. Finding individual word sense changes and their delay in appearance. In RANLP, pages 741–749, 2017.

[5] Noriaki Kawamae. Trend analysis model: trend consists of temporal words, topics, and timestamps. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 317–326. ACM, 2011.

[6] Qi He, Bi Chen, Jian Pei, Baojun Qiu, Prasenjit Mitra, and Lee Giles. Detecting topic evolution in scientific literature: how can citations help? In Proceedings of the 18th ACM conference on Information and knowledge management, pages 957–966. ACM, 2009.

[7] William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. arXiv preprint arXiv:1605.09096, 2016.

[8] John A Bullinaria and Joseph P Levy. Extracting semantic representations from word co-occurrence statistics: A computational study. Behavior research methods, 39(3):510–526, 2007.

[9] Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, and Suvrit Sra. Clustering on the unit hypersphere using von mises-fisher distributions. Journal of Machine Learning Research, 6:1345–1382, 2005.

[10] Janani Kalyanam, Amin Mantrach, Diego Saez-Trumper, Hossein Vahabi, and Gert Lanckriet. Leveraging social context for modeling topic evolution. In Proceedings of the 21th KDD, pages 517–526, 2015.

[11] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. Journal of machine Learning research, 3(Jan):993–1022, 2003.

[12] David M Blei and John D Lafferty. Dynamic topic models. In Proceedings of the 23rd ICML, pages 113–120, 2006.

[13] Jie Tang, Jing Zhang, Limin Yao, Juanzi Li, Li Zhang, and Zhong Su. Arnetminer: Extraction and mining of academic social networks. In Proceedings of the 14th KDD, pages 990–998, 2008.

[14] Thomas Hofmann. Probabilistic latent semantic analysis. In Proceedings of the 15th Conference on Uncertainty in Artificial Intelligence, pages 289–296, 1999.

[15] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In Advances in Neural Information Processing Systems, pages 3111–3119, 2013.

[16] Quoc V Le and Tomas Mikolov. Distributed representations of sentences and documents. In Proceedings of the 31st ICML, pages 1188–1196, 2014.

[17] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. Journal of Machine Learning Research, 9:2579–2605, Nov 2008.

[18] Ellen M Voorhees. Variations in relevance judgments and the measurement of retrieval effectiveness. Information Processing & Management, 36(5):697–716, 2000.

[19] Lawrence Hubert and Phipps Arabie. Comparing partitions. Journal of Classification, 2(1):193–218, 1985.

[20] David Newman, Jey Han Lau, Karl Grieser, and Timothy Baldwin. Automatic evaluation of topic coherence. In Proceedings of HLT-NAACL 2010, pages 100–108, 2010.