Exploring Concept Representations for Concept Drift Detection

Oliver Becher
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
becher@cwi.nl

Laura Hollink
Centrum Wiskunde & Informatica
Amsterdam, The Netherlands
hollink@cwi.nl

Desmond Elliott
University of Edinburgh
United Kingdom
d.elliott@ed.ac.uk

ABSTRACT

We present an approach to estimating concept drift in online news. Our method is to construct temporal concept vectors from topic-annotated news articles, and to correlate the distance between the temporal concept vectors with edits to the Wikipedia entries of the concepts. We find improvements in the correlation when we split the news articles based on the amount of articles mentioning a concept, instead of calendar-based units of time.

KEYWORDS

Concept drift, Vector representations, News, Wikipedia edits

1 INTRODUCTION

Concepts in Knowledge Organisation Systems (KOSs) are used to provide structured annotations and background knowledge in a wide variety of applications. They enhance interoperability between datasets and enable structured access to annotated document collections. These benefits, however, are compromised when concept change (or *drift*) occurs. Wang et al. [8] define three types of concept drift: (1) change in the intension of the concept, defined as the definition or the properties of the concept; (2) change in the extension, or the instances, of a concept; and (3) change in the label of the concept. Each type of concept drift may lead to problems for applications working with KOSs. For example, an annotation of a document may become invalid if the intension of the concept changes. Correspondences between two concepts in different KOSs may become incorrect if the extension of one of them changes. A user's keyword query on a historic corpus may be interpreted incorrectly if the (prevalent) label to refer to a concept has changed.

Significant progress has been made in the detection of meaning change of words (e.g., [3, 9]). They are based on distributional methods, where the meaning of a word is defined as the context in which it appears. A change in context over time may then signify a change in meaning. In this paper, we study change in the meaning of concepts in a KOS. Drawing inspiration from work on word-change detection, we aim to explore whether the change of a concept can be measured from changes in how it appears in the context of a document collection. This is different from other work on concept change in KOSs in the sense that we ignore changes in the structure of the KOS.

This paper is an initial step towards understanding how the context of a concept can be represented to effectively capture concept change. Our representation is based on the co-occurrence between concepts that appear as annotations of documents in a diachronic collection: if two concepts co-occur if they are annotations of the same document. Hence, a concept can be seen as a vector of co-occurrence counts with other concepts in the KOS. Concept change can then be measured by comparing vectors created for different time spans in the collection. We experiment with various versions of this basic idea, and apply it to detect change in an annotated document collection: the ION dataset of 300k online news articles, annotated with Wikipedia pages [4].

To evaluate our method, we use Wikipedia edit counts. This is based on the idea that a Wikipedia article is edited when a change to the page was needed; hence, a higher number of edits may signify a change in the underlying concept. Generally speaking, evaluation of concept drift detection methods is hampered by a lack of large scale evaluation datasets. Wikipedia edits are not to be seen as a gold standard of concept drift. While some edits might be due to a change in the concept, others might be, for example, additions of missing information or corrections of previous mistakes. Our assumption is that even though Wikipedia edit counts are a noisy signal with respect to concept change, a correlation between our change scores and the edits counts does say something about the effectiveness of our method.

2 REPRESENTING CONCEPTS

2.1 Creating Concept Vectors

Given a concept vocabulary C with N concepts, we create vector representations of the concepts through their usage in a document collection.

We assume there is a collection of time-ordered documents $\mathcal{D}$. A document d_i is annotated with M topic annotations $t_1, \ldots, t_M$, drawn from a total of T topics. Each document in the collection can be represented as a binary document topic vector, $\mathbf{d_i} \in \mathcal{R}^{1xT}$. An element in the document topic vector takes a value of 1 if the document has been annotated with that topic. We also assume a function f: T $\rightarrow$ C that maps between the topic annotation and concept vocabulary.

We construct a concept vector $\mathbf{c_j}$ for each concept in our vocabulary $c_1, \ldots, c_N$ from co-occurrence counts of the topic annotations in documents in the document collection. The set of concept vectors forms a sparse matrix $\mathbf{C} \in \mathcal{R}^{N \times N}$, where each row defines a concept through co-occurrence with other concepts.

SEMANTiCS 2017 workshop proceedings: Drift-a-LOD
September 11-14, 2017, Amsterdam, Netherlands

Our concept vectors are co-occurrence counts. We reduce the effect of frequently occurring concepts by re-weighting the vectors using a TF-IDF-like weighting scheme, so that tf-idf(c_i, c_j) = $tf(c_i, c_j) * idf(c_i)$, where $tf(c_i, c_j)$ is the number of times that concept c_i co-occurs with concept c_j and $idf(c_i) = log\frac{\text{N}}{df(C,c_i)} + 1$, with $df(C, c_i)$ as the count of c_i concept annotations in the entire concept vocabulary C.

2.2 Temporal Concept Vectors

Recall that we are interested in measuring the change in the meaning of a concept over time. We redefine C to include a temporal dimension, $\mathbf{V} \in \mathcal{R}^{NxNxK}$, where the third dimension represents K units of time, and $\sum_k \mathbf{V}_k = \mathbf{V} \in \mathcal{R}^{NxN}$. There are many ways to define K: the document collection can be split into days, weeks, months, or any other valid approach to splitting the collection according to the sequential ordering of the documents. Note that the co-occurrence statistics over topic annotations needs to be calculated such that only documents timestamped between consecutive units of time are used in the calculation, i.e. t=s_1 and t=s_2 are used to define a temporal concept vector $\mathbf{v_{j,s2}}$ at t=s_2.

2.3 Temporal Vector Distance

We measure the change in the meaning of concepts by comparing the vectors in the temporal concept matrix between subsequent units of time. Specifically, we measure the change in a concept $\mathbf{c_j}$ between time k and $k-1$ using a similarity metric $sim(\cdot, \cdot)$:

$$\text{distance}(\mathbf{v_j}, \text{s, s-1}) = sim(\mathbf{v_{j,s}}, \mathbf{v_{j,s-1}}) \quad (1)$$

We experiment with two similarity metrics: cosine similarity, previously used to detect concept drift [7], and KL-divergence (when the vectors represent distributions).

3 APPLICATION TO AN ANNOTATED NEWS COLLECTION

3.1 Dataset and Model Application

We explore our method for constructing concept representations and measuring concept change with a dataset of online news articles [4]. This data set contains news articles together with topic annotations and images in their natural textual context. The richness of information and meta data in this dataset can give many ways to define and explore concepts, while a defined structure of the data helps to use it reliably and consistently.

The dataset contains articles published online between August 2014 – August 2015. In total, it includes more than 300K articles from five publishers across British and US English sources: Daily Mail, The Independent, New York Times, Huffington Post, and the Washington Post. The articles are annotated with topics using TextRazor[1]. TextRazor uses Wikipedia as a topic vocabulary. This vocabulary ranges from narrowly defined concepts, e.g., The United States Women's Soccer Team or Electromagnetism, to broader concepts, e.g., Sport or Science. The average number of topic annotations per article is 25 broad 'Category pages' and 5 specific (non-category) pages, giving in total 122,000 distinct topic annotations on all articles.

[1] http://www.textrazor.com

We define our concept vocabulary C as a subset of TextRazor's topic vocabulary T: we retain only topics that are associated with at least 2 articles. In preliminary experiments, we found that concepts that are associated with too few articles have sparse representations resulting in unrealistic change scores between the representations. This leaves us with N=70,000 concepts. The mapping function f: T → C is trivial in this case. However, the structured nature of Wikipedia, and the links that it provides to other concept vocabularies, provide starting points for other mapping functions, allowing us to explore other concept vocabularies in the future.

We construct concept vectors using the method outlined in Section 2. The vocabulary of the concept vectors is defined over the Wikipedia entries, therefore it is trivial to map the topic annotations to the concept vectors.

3.2 Visualization

To visualize the change that a concept c has undergone, we create a stream graph [1] of the temporal concept vectors of c. Figure 1, for example, plots the temporal vectors of the Wikipedia concept Police. Each 'stream' represents a concept that co-occurs with Police in the document collection. The thickness of the line represents the co-occurrence count at a certain time period. Since stream graphs are suited to convey changes over time of only a limited number of concepts, we select only those that occur most frequently. Specificaly, we create 'streams' for only those concepts that are among the top 5 most frequently co-occurring concepts in any of the temporal concept vectors of concept c.

In figures 1, 2, and 3 we plot two concepts for which the average change is low (measured as a high average cosine similarity between 12 temporal vectors) and one where the average change score is high. Figure 1 shows that Police is a stable concept: the top five most frequently occurring concepts remain frequent troughout the year, and the volume of documents in which they co-occur hardly fluctuates. However, a concept might change on a larger time scale than given in the data. Nonetheless, Police seems to be more stable than other concepts in the time span.

The concept Labour_Party (Figure 2) is stable as well: although there is a burst in the volume of documents about this concept, there is hardly a change in which concepts co-occur in these documents. In other words, there is change in *how much* reporting there is about the Labour_Party, but not in *how* they are reported.

Figure 3 shows the streamgraph of the New York University. We can see that the most co-occurring topics are constantly changing in the streamgraph, both in periods with a high volume of documents and in periods with a low volume of documents. This suggests changes in *how much* and *how* New York University has been reported in the news.

4 TOWARDS A QUANTITATIVE EVALUATION

4.1 Measuring Concept Change

Concept change detection is hard to evaluate for a lack of gold standard datasets [5]. Kenter et al. [6] use a small sets of 21 human-judged change scores. Frermann and Lapata [2] indirectly evaluate change detection by using it in an application for which a gold standard exists, namely the SemEval task for dating text. To the best

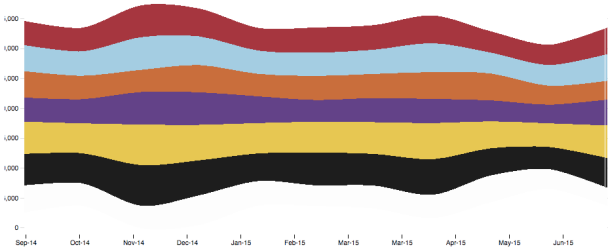


Figure 1: WP:Police streamgraph shows a stable set of top-5 concepts in its temporal vectors. (See Section 3.2 for more details.)

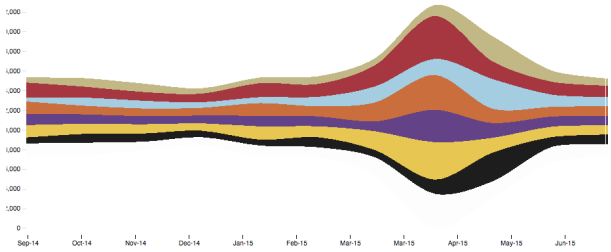


Figure 2: WP:Labour_Party_(UK) streamgraph has a stable set of top-5 concepts but a lot of activity centered around a specific time.

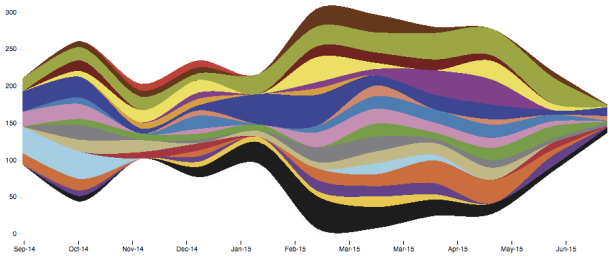


Figure 3: WP:New_York_University streamgraph undergoes constantly shifting concept representation in our dataset.

of our knowledge, large scale datasets to directly valuate change detection, do not exist.

For our application, we explore the use of Wikipedia edit rates to evaluate our method of concept change representation. We believe that the act of editing a Wikipedia page can signal a change in the information that is relevant to that entry.

Specifically, given a concept c and a pre-defined K units of time, we measure changes scores as the consecutive temporal vector distances for concept c (Section 2.3); then, we count the number of Wikipedia edits to the aligned article during each of the K units of time. We evaluate our method by measuring the Spearman correlation between the change scores (i.e. the temporal vector distances) and the Wikipedia edit counts. The higher the correlation, the more accurately the temporal concept vectors can estimate the rate of change of the Wikipedia entries.

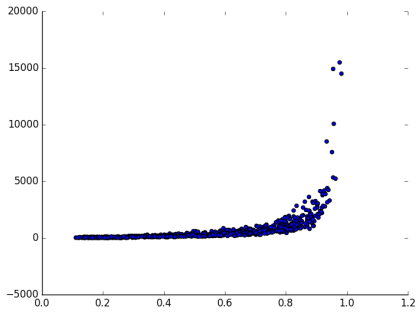


Figure 4: Scatter plot of average cosine similarities and annotation count of all concepts

We perform an experiment on 964 concepts. Since it seems likely that the number of articles that a concept is related to plays a role, we draw a stratified random sample from our concept vocabulary to include both frequently and infrequently used concepts. We select three different strati of even size. Group 1 contains concepts which are related to more than 500 articles. Group 2 contains concepts which are related to at least 200 articles but not more than 500. Group 3 contains concepts with at least 24 articles but less than 200. The sample includes only concepts that map to 'regular' Wikipedia pages and not Category pages.

Figure 4 plots the number of articles that a concept is related to against the average cosine similarity between the temporal vectors of that concept. This shows that the more frequent a concept is used as an annotation, the higher the average cosine similarity, i.e., the lower the change. This is analogous to the change of meaning of words [3], where the semantic changes of words scale with inverse frequency, known as the *law of conformity*.

We compare four models, each with different settings regarding the way that time units are set, the use of TF-IDF, and the choice of similarity measure (either cosine similarity or KL-divergence).

4.2 Models

4.2.1 Fixed Time Bins (Cosine). Starting with the most basic setup of our method, we calculate temporal concept vectors for time frames (or *bins*) of a fixed duration. With n time frames, each frame covers an $n/year^{\text{th}}$ of the dataset. For example, with 52 frames, each frame covers exactly one week. We use the cosine similarity to calculate change scores between each temporal concept vector.

4.2.2 Flexible Time Bins (Cosine). In this model, we calculate temporal concept vectors for time periods that each cover a fixed amount of articles. Thus, time frames differ in length of days rather than amount of data. The amount of articles per bin depends on the total amount of articles available per concept. Analogous to Fixed Time Bins, we create n bins, therefore we assign a nth of the total amount of articles to each bin. However, a concept may have such an amount of articles that does not split evenly into n bins. Thus, it may be split into more than n bins. With these vectors, we use the cosine similarity to calculate change scores. We use the same time frames to bin the Wikipedia edits and estimate a correlation.

Run / n bins	> 100	> 52	> 24	> 12	> 6
Fixed Time Bins (Cos)	0.07	0.18	0.22	0.36	0.26
Flexible Time Bins (Cos)	-0.2	-0.19	-0.14	0.03	0.33
- TF-IDF (Cos)	-0.2	-0.2	-0.13	0.0	0.26
Flexible Time Bins (KL)	0.23	0.25	0.29	0.19	-0.3

Table 1: Average Spearman correlation between concept similarity scores and Wikipedia edits. Negative correlations are good for cosine similarity; positive correlations are good for KL-divergence.

4.2.3 Flexible Time Bins (No TF-IDF, Cosine). Exactly the same as Flexible Time Bins (Cosine) except we do not re-weight the temporal concept vectors using TF-IDF.

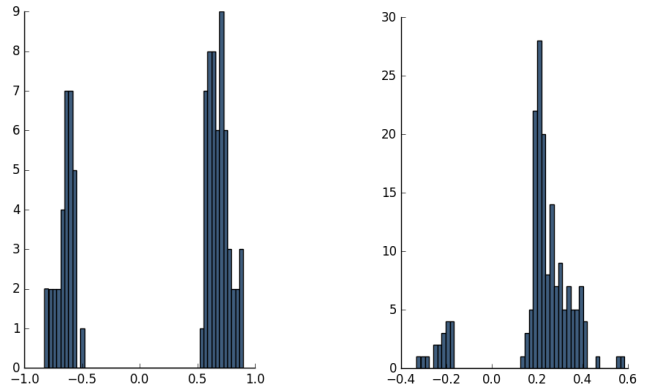
4.2.4 Flexible Time Bins (KL-divergence). This model is identical to Flexible Time Bins except we measure the distance between temporal concept vectors using Kullback-Leibner divergence (KL) instead of cosine similarity.

4.3 Results

We collect Spearman correlation coefficients for 964 concepts using different numbers of time frames (6, 12, 24, 52, and 100). Table 1 shows the average correlation over concepts that are significantly correlated with Wikipedia edits. Note that the experiments with Cosine similarity measure between temporal concepts should return a negative correlation, while the experiments with the KL-divergence distance should return a positive correlation. The results in Table 1 show that the performance of the models decreases as we decrease the number of time bins.

The Fixed Time Bins (Cosine) model only returns positive correlations, indicating that fixed units of time (in this case, splitting the articles into months) does not act as a reliable proxy for concept change in our dataset. The Flexible Time Bin experiments (Cosine) and (-TF-IDF) are better correlated with Wikipedia edits than the Fixed Time Bin model. We do not find a difference in not re-weighting the concept vectors using TF-IDF. Finally, we find a small improvement from using KL-divergence as the temporal vector distance metric instead of Cosine similarity. Throughout, we can see that the number of temporal bins n is a crucial parameter in our experiment.

We performed a follow-up analysis of the effect of the number of temporal bins. The histograms in Figures 5a to 5b show the distributions of the Spearman correlations for the Flexible Time Bins (KL) model with n=12 or n=100. We find that the ratio of positively correlations to negative correlations is substantially reduced by having more time bins. More time bins clearly improves the quality of the concept vectors.



(a) n=12 **(b) n=100**

Figure 5: Distribution of Spearman correlations for Flexible Time Bins (KL) with n=12 (left) or n=100 (right) time bins. The ratio of positive/negative correlations is much improved by having more time bins.

5 CONCLUSION AND FUTURE WORK

We explored concept change using vector space concept representations. The concept vectors were constructed from topic co-occurrence in a large collection of online news articles. We introduced a temporal aspect to the vectors by requiring the co-occurrences to happen within pre-defined windows of time. We explored to what extend concept change can be evaluated by correlating the distance between its temporal concept vectors and edits to the Wikipedia article corresponding to the concept.

We found that a flexible approach to defining a window of time was more successful than using calendar-based windows of time. We also found that having more windows of time resulted in better correlations between the temporal vector distances and Wikipedia article edits.

Future work includes an analysis of which types of concepts correlate to Wikipedia edits counts, to get more insights into the use of Wikipedia as an evaluation tool. Similarly, we could look into the types of edits made on Wikipedia to distinguish actual change from simple growth of an article.

REFERENCES

[1] Lee Byron and Martin Wattenberg. 2008. Stacked Graphs - Geometry and Aesthetics. *IEEE transactions on visualization and computer graphics* 14 6 (2008), 1245–52.

[2] Lea Frermann and Mirella Lapata. 2016. A Bayesian Model of Diachronic Meaning Change. *Transactions of the ACL* 4 (2016), 31–45.

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

[4] Laura Hollink, Adriatik Bedjeti, Martin van Harmelen, and Desmond Elliott. 2016. A Corpus of Images and Text in Online News. (2016).

[5] Laura Hollink, Sándor Darányi, Albert Meroño Peñuela, and Efstratios Kontopoulos. 2017. First Workshop on Detection, Representation and Management of Concept Drift in Linked Open Data: Report of the Drift-a-LOD2016 Workshop: Front Matter.. In *Knowledge Engineering and Knowledge Management. EKAW 2016 (Lecture Notes in Computer Science)*, Vol. 10180. 15–18.

[6] Tom Kenter, Melvin Wevers, Pim Huijnen, and Maarten de Rijke. 2015. Ad hoc monitoring of vocabulary shifts over time. In *Proceedings of the 24th International Conference on Information and Knowledge Management*. 1191–1200.

[7] Astrid van Aggelen, Laura Hollink, and Jacco van Ossenbruggen. 2016. Combining distributional semantics and structured data to study lexical change. In *European Knowledge Acquisition Workshop*. Springer, 40–49.

[8] Shenghui Wang, Stefan Schlobach, and Michel Klein. 2011. Concept drift and how to identify it. *Web Semantics: Science, Services and Agents on the World Wide Web* 9, 3 (2011), 247 – 265. https://doi.org/10.1016/j.websem.2011.05.003 Semantic Web Dynamics Semantic Web Challenge, 2010.

[9] Yating Zhang, Adam Jatowt, and Katsumi Tanaka. 2016. Towards understanding word embeddings: Automatically explaining similarity of terms. *2016 IEEE International Conference on Big Data (Big Data)* (2016), 823–832.