Non-Temporal Orderings as Proxies for Extensional Concept Drift

Albert Meroño-Peñuela[1,2] and Stefan Schlobach[1]

[1] Department of Computer Science, VU University Amsterdam, NL
albert.merono@vu.nl
[2] Data Archiving and Networked Services, KNAW, NL

Abstract. In census data, *concepts* are central entities represented by variables and their values. The meaning of these concepts is often assumed to be stable, but in fact it can change over time: we call this *concept drift*. Extensional concept drift is one type of change of meaning that affects the things the concept extends to, having drastic consequences on longitudinal querying. In this paper we detect extensionally drifted concepts in current Linked Census Data when a time ordering of such concepts is not available. We exploit the Linked Data cloud to obtain meaningful proxies for such orderings.

Keywords: Concept Drift, Semantic Web, Linked Census Data

1 Introduction

Most linked datasets assume some degree of stability in the concepts (variables, values) they refer to. But the meaning of these concepts can change over time. In this paper we find and report back this change of meaning of concepts, or concept drift, in two census datasets. Concept drift can happen at the concept identifier level (*label drift*), in the basic properties of the concept (*intensional drift*), or to the things the concept refers to (*extensional drift*) [9]. This paper proposes a statistics-based solution for the latter.

Concept drift is often assumed to happen between two time gapped variants of a concept. Hence, *time* is the fundamental ordering of concepts in which concept drift occurs. But time series are not available for the datasets we work with. In this paper we propose a set of concept orderings that do not include time, and we show their usefulness as proxies for concept drift detection. To get such orderings, we exploit Linked Data to enrich and complement the census data we already have.

This paper is organised as follows. In Section 2 we describe the state of the art in concept drift. In Section 3 we set the formal framework for the study of concept drift. In Section 4 we describe experiments to detect extensional concept drift in the Australian and French censuses in the absence of time series. Finally, in Section 5 we establish some conclusions.

2 Related Work

In Machine Learning, concept drift is defined as the situation in which the statistical properties of a target variable change over time in unforeseen ways [8]. Several concept drift detection algorithms have been developed in this setting [2,4,6]. On the Semantic Web, concept drift relates to the study of the dynamics of meaning. This has been addressed in the field of ontology change and evolution [1], in Description Logics [3], and in knowledge management [9].

3 Concept Drift

As reality changes continuously, concepts also change over time. A concept refers to different objects, real or abstract, at different points in time. We use the formalisation framework described by Wang et al. [9] in order to study concept drift over time.

Definition 1. *The* meaning of a concept *C is a triple (label(C),int(C),ext(C)), where label(C) is a string, int(C) a set of properties (the* intension *of C), and ext(C) a subset of the universe (the* extension *of C).*

All the elements of the meaning of a concept can change. To address concept identity over time, Wang et al. [9] assume that the intension of a concept C is the disjoint union of a rigid and a non-rigid set of properties (i.e. $(int_r(C) \cup int_{nr}(C))$). Then, a concept is uniquely identified by some essential properties that do not change. The notion of identity allows the comparison of two variants of a concept at different points in time, even if a change on its meaning occurs.

Definition 2. *Two concepts C_1 and C_2 are considered identical if and only if, their rigid intension are equivalent,* i.e.*, $int_r(C_1) = int_r(C_2)$.*

If two variants of a concept at two different times have the same meaning, there is no concept drift. We define intensional, extensional, and label similarity functions $sim_{int}, sim_{ext}, sim_{label}$ to quantify meaning similarity. Each of these functions has range $[0, 1]$, and a similarity value of 1 indicates equality.

Definition 3. *A concept has extensionally drifted in two of its variants C' and C", if and only if, $sim_{ext}(C', C'') \neq 1$. Intensional and label drift are defined similarly.*

To apply this framework of concept drift it is required to define intension, extension and labelling functions, and to define similarity functions over intension, extension and labels. We define these functions in Section 4.2.

4 Meaningful Orderings as Concept Drift Proxies

In this section we apply the concept drift framework presented in Section 3 to study the change of meaning of concepts in RDF Data Cube versions of the Australian census of 2011 and the French census of 2010.[3,4,5] More concretely, we apply the notion of *extensional drift* to detect extensionally drifted concepts in these censuses. Concept drift is usually assumed to happen between two time gapped variants of a concept. Hence, *time* is the fundamental dimension to order such variants. Since time series are not available for these datasets, in this paper we propose a different set of concept orderings, and we study their applicability. To get such orderings, we exploit Linked Data to complement the census data we already have.

4.1 Data Retrieval

We query the Australian and French census datasets from the statistical environment R [7] via the SPARQL R package [5].[6] We select the variables gender, age range, location, labour status and population. In the Australian census we query data at the state level, and in the French census we aggregate results at the *departement* level.

To extend these variables we query DBPedia[7]. In the Australian case, we ask for the *gross domestic product (GDP) per capita* of all states. In the French case, we ask for the area and total population of all *departements*, and we derive the *population density* for each of them.

4.2 Non-Temporal Extensional Concept Drift

We are interested in detecting extensional concept drift, that is, $sim_{ext}(C', C'') \neq 1$ for two given variants C', C'' of a concept C (see Section 3). Intuitively, this means that the instances of C have changed significantly. We interpret extensional concept drift in a statistical setting. We define the extension function $ext(C)$ as the function that returns the number of individuals that belong to C, and the extension similarity function $sim_{ext}(C', C'')$ as the function that returns the probablity that C' and C'' have identical populations. We assume that the extension of C has drifted between C' and C'' iff the populations of C' and C'' are non identical (there is a shift between the populations of C' and C'').

We choose the concept of *youth unemployment* to study its extensional drift in both censuses. To replace the natural ordering of time in the occurrences of

[3] See `http://www.datalift.org/en/event/semstats2013/challenge`

[4] SPARQL endpoint serving the datasets at `http://lod.cedar-project.nl:8080/sparql/semstats/`

[5] Source code at `https://github.com/albertmeronyo/ConceptDrift/blob/master/stats/semstats-challenge.R`

[6] SPARQL queries at `https://github.com/albertmeronyo/ConceptDrift/blob/master/sparql/semstats-challenge.txt`

[7] `http://dbpedia.org/sparql`

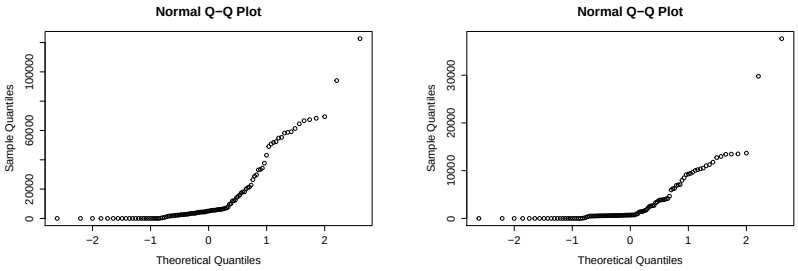


Fig. 1: Normal QQ-plots of all population counts in Western Australia and Tasmania. Both plots reveal non-normality of their distributions.

```
> wilcox.test(x,y)

        Wilcoxon rank sum test with continuity correction

data:  x and y
W = 16, p-value = 0.02857
alternative hypothesis: true location shift is not equal to 0
```

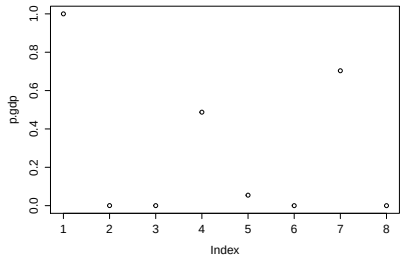
Listing 1.1: Wilcoxon test for the population counts of unemployed young people in Western Australia and Tasmania

this concept, we use the variables *GDP per capita* of the Australian states and *population density* of the French *departements* to order such occurrences.

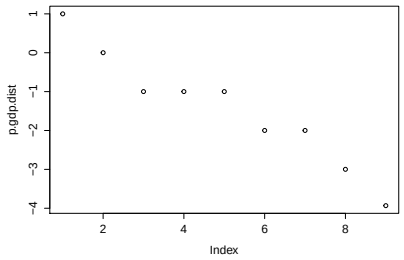
As an example, we calculate the extensional drift of youth unemployment in the Australian states of Western Autralia and Tasmania (highest and lowest GDP per capita, respectively). We want to know if population counts of unemployed young people (15-24 years old) have identical data distributions between these regions. Without assuming the data to have normal distribution (see Figure 1), we want to test at .05 significance level if the population counts for youth unemployment have identical data distributions.

The null hypothesis, H_0, is that the young unemployed people from these two regions are identical populations. To test the hypothesis, we run the Wilcoxon signed-rank test that comes with the R distribution [10]. We run the `wilcox.test` function using these samples (see Listing 1.1), concluding that the population of unemployed people between 15 and 24 in Western Australia and Tasmania are statistically non-identical populations ($p < 0.05, N = 4$, Wilcoxon signed-rank test). Consequently, there is extensional drift in this case.

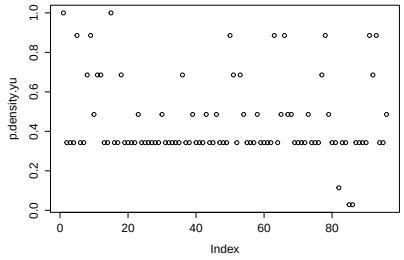
In order to have a complete overview on how youth unemployment evolves as GDP per capita increases, we run the same test for all Australian region pairs, in GDP per capita ascending order. The resulting p-values indicate whether there is an extensional drift between the regions ($p < 0.05$, see Figure 2) or, on the contrary, the concept remains stable. To view the evolution of extensional drift on a relative scale, for each drift test k we compute the distance function



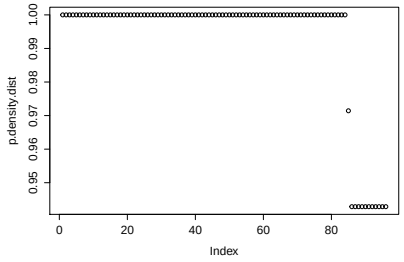
(a) Extensional drift per Australian region. P-values below 0.05 denote drift. Regions by ascending GDP per capita.



(b) Evolution of relative distances d_k in Australian regions. A decrease in y-values denotes drift.



(c) Extensional drift per French region. P-values below 0.05 denote drift. Regions by ascending population density.



(d) Evolution of relative distances d_k in French regions. A decrease in y-values denotes drift.

Fig. 2: Plots of p-values and d_k distances for extensional drift in Australian and French regions for the concept *youth unemployment*. The x-axis represents ascending regions per GDP per capita (a, b) and population density (c, d); the y-axis represents p-values (a, c) and relative drift distances d_k (b, d).

$$d_k = \begin{cases} p_{k-1} - \alpha(p_k) & \text{if } p_k < 0.05 \\ p_{k-1} & \text{if } p_k \geq 0.05 \end{cases}$$

where α is a function that magnifies the distances in case of drift (Figure 2).

We evaluate the applicability of this method with different data and ordering criteria. We repeat the *youth unemployment* experiment, this time on the French census. We use *population density* of the *departements* as the ordering to compare different variants of the same concept. Results are also shown in Figure 2.

5 Conclusions

Figure 2 shows meaningful results on the use of GDP per capita and population density to track the evolution of the extensional drift of youth unemployment.

In the Australian case, the population distributions tend to vary in the less rich regions, and they stabilize as the regions get richer. The top two regions also differentiate themselves from the rest. In the French case, there is a great stability of the distributions until a drastic change happens when approaching the top 20% richest regions, which probably reveals differences in how these labour markets behave. We consider our selected orderings to be as meaningful and useful as time for the applicability of our extensional concept drift detection method.

In this paper we present the application of an extensional concept drift detection method in Linked Census Data when temporal variants of the concepts are not available. Concretely, we study extensional drifts of the concept *youth unemployment* in the Australian and French censuses, leveraging Linked Data to retrieve meaningful orderings of the data in the absence of temporal orderings.

Acknowledgements The work on which this paper is based has been partly supported by the Computational Humanities Programme of the Royal Netherlands Academy of Arts and Sciences, under the auspices of the CEDAR project. For further information, see http://ehumanities.nl. This work has been supported as well by the Dutch national program COMMIT.

References

1. Fanizzi, N., d'Amato, C., Esposito, F.: Conceptual Clustering: Concept Formation, Drift and Novelty Detection. In: The Semantic Web: Research and Applications, 5th European Semantic Web Conference. LNCS 5021. pp. 318–332. Springer (2008)
2. Flouris, G., Manakanatas, D., Kondylakis, H., Plexousakis, D., Antoniou, G.: Ontology change: classification and survey. The Knowledge Engineering Review 23(2), 117–152 (2008)
3. Gonçalves, R.S., Parsia, B., Sattler, U.: Analysing Multiple Versions of an Ontology : A Study of the NCI Thesaurus. In: Proceedings of the 24th International Workshop on Description Logics (DL 2011). vol. 745. CEUR Workshop Proceedings (2011), http://ceur-ws.org/Vol-745/
4. Gulla, J.A., Solskinnsbakk, G., Myrseth, P., Haderlein, V., Cerrato, O.: Semantic Drift in Ontologies. In: Proceedings of the 6th International Conference on Web Information Systems and Technologies. vol. 2. INSTICC Press (2010)
5. van Hage, W.R., with contributions from: Tomi Kauppinen, Graeler, B., Davis, C., Hoeksema, J., Ruttenberg, A., Bahls., D.: SPARQL: SPARQL client (2013), http://CRAN.R-project.org/package=SPARQL, R package version 1.15
6. Klein, M.: Change Management for Distributed Ontologies. Ph.D. thesis, VU University Amsterdam (2004)
7. R Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2013), http://www.R-project.org/
8. Tsymbal, A.: The problem of concept drift: definitions and related work. Tech. Rep. TCD-CS-2004-15, Computer Science Department, Trinity College Dublin (2004)
9. Wang, S., Schlobach, S., Klein, M.C.A.: What Is Concept Drift and How to Measure It? In: Knowledge Engineering and Management by the Masses - 17th International Conference, EKAW 2010. Proceedings. pp. 241–256. Lecutre Notes in Computer Science, 6317, Springer (2010)
10. Wilcoxon, F.: Individual comparisons by ranking methods. Biometrics Bulletin 1(6), 80–83 (1945)