A Study of Intensional Concept Drift in Trending DBpedia Concepts

Albert Meroño-Peñuela
Department of Computer Science
Vrije Universiteit Amsterdam
Amsterdam, The Netherlands
albert.merono@vu.nl

Efstratios Kontopoulos
Information Technologies Institute
Thessaloniki, Greece
skontopo@iti.gr

Sándor Darányi
Swedish School of Library and Information Science
University of Borås
Borås, Sweden
sandor.daranyi@hb.se

Ioannis Kompatsiaris
Information Technologies Institute
Thessaloniki, Greece
ikom@iti.gr

ABSTRACT

Concept drift refers to the phenomenon that concepts change their intensional composition, and therefore meaning, over time. It is a manifestation of content dynamics, and an important problem with regard to access and scalability in the Web of Data. Such drifts go back to contextual influences due to social embedding as suggested by e.g. topic analysis, news detection, and trends in social networks. Using DBpedia as a source of timestamped Linked Open Data, we analyze the interaction between a sample of popular keywords, as recorded by Google Trends, and their respective concept drifts in DBpedia. For the latter task, we deploy SemaDrift, an ontology evolution platform for detecting and measuring content dislocation dependent on context modification. Our hypothesis is that social embedding and awareness is an important trigger for concept drift in crowdsourced knowledge bases on the Web.

KEYWORDS

Concept Drift, Semantic Web, DBpedia, Wikipedia, Google Trends

1 INTRODUCTION

Rather than remaining stable, permanent, and fixed, the meaning of concepts changes over time. The *Historical Thesaurus of the Oxford Dictionary of English*[1] shows how definitions attributed to words are different in different periods of history. In the Dutch historical censuses (1795-1971) [15] the taxonomy of occupations shows an extraordinary variation every decade, in line with the major transformations of labor in the society of that time. We call the change of meaning of concepts over time *concept drift*. Concept drift can have drastic effects in the performance of a system, like changing queries and inconsistent analyses.

What *causes* concept drift to occur in these systems? In the specific setting of the Semantic Web [3] (now also referred to as Web of Data), concepts in ontologies and taxonomies are regularly updated by humans in order to “reflect changes in the real world, changes in user requirements, and drawbacks in the initial design” [23]. Hence, concept drift in semantic systems has a traceable and direct origin in humans. However, the more recent trend on Linked Data [9] in the Semantic Web, rather than manually building these ontologies and taxonomies, has automated the way in which semantic systems obtain their concepts. A canonical example is DBpedia [14], which relies largely on automated knowledge extraction methods to create Linked Data out of Wikipedia[2]. In this situation, the causes of concept drift become more difficult to trace.

There are various plausible explanations for the origin of concept drift in complex systems. One of them is the interaction of evolving context with evolving content. Social awareness (instigated by events or the media) triggers a process of knowledge sharing on the Web. This process often results in changes in knowledge bases, which may have an impact in the meaning of concepts. Wikipedia, the biggest collaboratively-built knowledge base of the Web, has been criticized for “allegedly exhibiting systemic bias, presenting a mixture of truths, half truths, and some falsehoods, and, in *controversial topics*, being subject to manipulation and spin” [18]. It is then worth considering whether the controversy, novelty or burst of a topic has an impact on how reality is formally defined in knowledge bases derived from Wikipedia, such as DBpedia [14].

In this paper we propose a framework to measure the influence of user engagement on the Web with its effects on concept drift in Web crowd-sourced databases. We are interested in the process of public

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

[1] http://public.oed.com/historical-thesaurus-of-the-oed/
[2] https://www.wikipedia.org/

opinion influencing the feature composition of concepts as captured by automatic means. Hence, our research question is: *what patterns of influence can we discern between trends in queries by Web users, and concept drift in crowd-sourced databases?* To address this question, we propose a tool chain that quantifies the *trendiness* of Web queries, and confronts it with *measures of concept drift* for Linked Data. This tool chain consists of SemaDrift [20], a concept drift measuring platform; Google Trends[3], an index of the popularity of Web user queries over time; and the different versions of DBpedia accessible via Linked Data Fragments (LDF) [27].

Concretely, the contributions of this paper are:

- An automated and systematic way for retrieving time-specific concept intensions from Linked Data sources (Section 3.1);
- A framework for studying the relationship between the *popularity of Web user queries* and the *drift of their associated concepts* over time (Section 3);
- An experimental application of this framework to recent Web trending queries and the latest snapshots of DBpedia (Section 4).

2 RELATED WORK

The problems of semantic change and drift concern various research fields. In the areas of Semantic Web and knowledge representation, ontology evolution [13] addresses "the timely adaptation of an ontology and consistent propagation of changes to dependent artifacts" [1]. Features of evolution have been studied [22] and used for prediction using machine learning [17]. Gonçalves et al. [7] use Description Logics to calculate differences between ontologies (so-called *semantic diffs*). Wang et al. [28] define the semantics of concept change and drift, and how to identify them. General surveys of semantic change in other fields, including language, have recently appeared [20]. On the use of trends of Web user queries and changing semantics, the work by Tiddi et al. [26] illustrates the use of knowledge from the Semantic Web to explain patterns in data, in particular on finding *causes* for trending queries in Google Trends. To the best of our knowledge, no previous work addresses the cause-effect relationship between trends and concept drift.

Standard means of observing changes in content include e.g. recognizing news in texts by topic detection and tracking [2], and new event or burst detection [16], which are in essence similar to time series analysis. Significant solutions range from extracting time-varying features from texts [24] to constructing timelines for event classification based on word usage statistics [25] and personalized newsfeeds based on information novelty [6]. In the latter, the inter- and intra-document dynamics of documents is considered to model how information evolves over time from article to article, as well as within individual articles. Such methods can be applied to the analysis of temporal dynamics in online text streams such as newsfeed or e-mail [11, 12], or chronologically ordered documents [5]. These are models typically based on graph theory vs. vector space methods vs. probability theory, capturing local vs. global context of content as a basis of the results, therefore our current models of content are context-dependent. However, this dependency, although acknowledged, is typically not quantified, a precondition for improved models.

In general terms, another relevant track is research into time series of content. From a Natural Language Processing (NLP) perspective, a typical example is to study *diachronic collocations*: a word's company (its collocates) may change over time, reflecting changes in that word's meaning and/or in the focus of the discourse in which it is embedded. However, traditional collocation extractors treat the underlying text corpus as a homogenous whole, and thus cannot adequately account for such diachronic changes in a word's collocation behavior, hence the need for a combination of diachrony and contextuality [10]. From an information science perspective, the study of *conceptual dynamics* [4] offers another comprehensive set of considerations. By the mathematical models they exploit, both tracks preserve the underlying contextual dependency of word content or meaning, ultimately going back to Harris' distributional hypothesis [8].

3 TRENDING CONCEPTS AND CONCEPT DRIFT

In this section we describe a workflow for studying the relationship between the popularity of Web user queries and the drift in concepts contained therein:

(1) We use an extended **LDF client** to systematically retrieve time-specific concept *intensions* of a chosen concept C (see Section 3.2) from compatible Linked Data sources with the Linked Data Fragments backend[4];
(2) Using the concept intensions retrieved in the previous step, we use **SemaDrift** [19] to measure intensional concept drift over $int(C)$. This represents how much the concept C has drifted in a certain time period;
(3) Finally, we **confront values of Trend and Drift**, and we observe the relationship between measurements of concept drift for the concept C, and measurements of popularity for a Web user query $q(C)$ that matches C.

3.1 Temporal DBpedia Concepts with LDF

Wikipedia is "a free online encyclopedia with the aim to allow anyone to edit articles"[5]. Aligning with the mission of Linked Data and the Semantic Web, DBpedia [14] aims at extracting structured content from Wikipedia, providing a means for semantically querying relationships and properties of its content. We assume this structured content of DBpedia resources to formally represent the meaning of their associated concepts. In this first step, we select a concept of interest C, and we query DBpedia to get the intension of C, $int(C)$ (i.e. its defining properties), at various points in time.

Querying massive Linked Data sources like DBpedia entails various challenges. One approach includes submitting processing-intensive queries to SPARQL endpoints; another approach is to download and locally query massive data dumps that are possibly not up-to-date. *Linked Data Fragments* (*LDF*) provide a conceptual framework that delivers a uniform view on RDF interfaces, aiming to minimize server resource usage while still enabling clients

[3] https://trends.google.com/

[4] To the best of our knowledge, currently DBpedia is the only Linked Data source with such support.

[5] https://en.wikipedia.org/wiki/Wikipedia

to query data sources efficiently [27]. In this work, we have deployed an openly available Java LDF client[6], which we extended for measuring intensional drift via the SemaDrift API.

3.2 Concept Drift and SemaDrift

To measure concept change between two versions of an ontology, we use the concept drift framework proposed by Wang et al. [28], which quantifies the change of meaning of concepts over time. In this framework, the *meaning of a concept* C is defined as the combination of its *intension*, *extension*, and *label*. The intension of C, $int(C)$, is the set of formal, explicit properties that axiomatically define C. The extension of C, $ext(C)$, is the set of its instances. The label of C, $label(C)$, is a human-readable string representing C.

Over time, $int(C)$, $ext(C)$ and $label(C)$ can change, and compromise the identity and traceability of C. To address this, the framework assumes that $int(C)$ is the disjoint union of rigid and non-rigid sets of properties, $int(C) = int_r(C) \cup int_{nr}(C)$). $int_r(C)$ uniquely identifies C by some essential properties that do not change. This allows the comparison of two variants of a concept at different points in time, even if $int_{nr}(C)$, $ext(C)$ or $label(C)$ change.

If two variants of C at two different times have identical $int(C)$, $ext(C)$ and $label(C)$, then there is no concept drift. Otherwise, the framework defines intensional, extensional, and label similarity functions $sim_{int} \mapsto [0,1], sim_{ext} \mapsto [0,1], sim_{label} \mapsto [0,1]$ to quantify meaning similarity. Then, there is extensional (intensional, label) *concept change* between two variants of C, C' and C'', iff $sim_{ext}(C', C'') \neq 1$.

Using the above definitions as its foundation, SemaDrift [20] constitutes a cutting edge suite of metrics and tools for measuring concept drift in different versions of an ontology, under an ontology evolution perspective. As demonstrated in [21], SemaDrift is totally domain agnostic, offering the capability of applying the underlying metrics and methods to any ontology originating from any domain of application. The platform consists of (a) an API for programmatically accessing the core drift measuring methods, (b) a Protégé plug-in [19], and, (c) a standalone desktop application. The full suite is available at http://mklab.iti.gr/project/semadrift-measure-semantic-drift-ontologies.

In this work we are deploying the core SemaDrift API, and we are particularly monitoring intensional drifts of DBpedia concepts; i.e. each DBpedia entry is essentially a class instance with associated properties, thus it makes no sense to measure drifts in its extension (instances have no extension) or label (entries in DBpedia maintain their labels unaltered).

3.3 Confronting Trends with Drift

Google Trends (GT) is a Web service that shows how often a particular search-term is entered relative to the total search-volume of the Google Search engine. For example, it is possible to compare the relative volume of queries between the search terms *Donald Trump* and *climate change* in a certain time period. These relative volumes of search-terms are given with a measurement from 0 (no volume) to 100 (maximum volume). In order to obtain these, a matching needs to be made between the chosen concept of interest C and its corresponding search-term query, $q(C)$, which is not trivial. For example, the DBpedia concept *Terrorism in the European Union* only matches the search-term *terrorism in europe* in GT. In this workflow we align C and $q(C)$ manually. Next, we normalize the GT scores by picking a comparatively popular and stable topic over time that sets the maximum score (e.g. *iPhone*).[7] All subsequent trend scores for other concepts are relative to this reference concept. We define the GT score for a concept C at time t as $GT(C, t)$. Finally, we define the two proxies of *popularity* and *trendiness* of a concept, $p(C)$, $t(C)$, as the arithmetic mean and standard deviation over the GT scores, respectively:

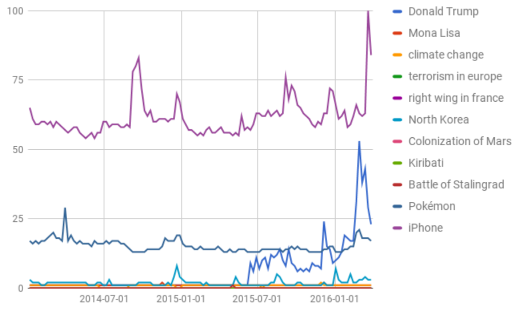

Figure 1: Chosen concepts and GT scores (2014-01 – 2016-04).

$$p(C) = \frac{1}{n} \sum GT(C,t), \;\; t(C) = \sqrt{\frac{1}{n} \sum (GT(C,t) - p(C))^2}$$

4 PRELIMINARY EVALUATION

In order to evaluate our framework, we propose a preliminary experiment to *measure the relationship of Web user queries in the intensional concept drift of DBpedia concepts between January of 2014 and April of 2016*. By this we adapt Harris' distributional hypothesis to RDF statements, i.e. we assume that intensional concept drifts go back to the social embedding of the detection environment, in other words, the feature composition of concepts is context-dependent.

To do so, we sample a small ($N = 11$) set of DBpedia concepts C and their equivalent search-terms $q(C)$ GT scores on that period. The chosen concepts, together with their GT scores over time, are shown in Figure 1. We chose these concepts considering one interest group, with both trendy and popular concepts (*iPhone*, *Donald Trump*, *Pokemon*); and a control group, with concepts of scarce trendiness and popularity (*Mona Lisa*, *Colonization of Mars*, *Battle of Stalingrad*).

4.1 Results

We use SemaDrift to calculate the intensional concept drift values of $int(C)$ for the chosen set of concepts of Figure 1[8]. Figure 2 confronts these intensional concept drift values with their popularity/trendiness $p(C)$, $t(C)$ scores derived from GT.

In Figure 2 we can observe an expected distribution over the x-axis of non-trendy vs. trendy concepts, to the left and the right, respectively. However, the patterns of intensional concept drift with respect to variations in trends are not as expected. Quite

[6] https://github.com/LinkedDataFragments/Client.Java

[7] We do this by using GT's *Most searched* feature over matching time periods.

[8] A detailed table with all drifting values and relevant predicates can be found at https://goo.gl/yQ531r.

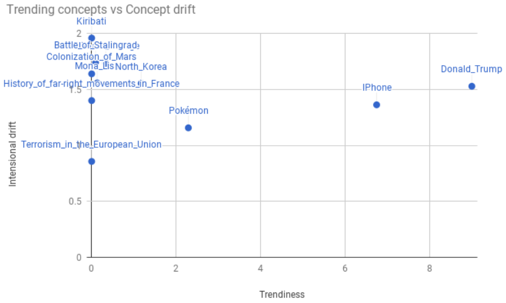

Figure 2: Trendiness *vs.* intensional concept drift.

the contrary: *the highest concept drift measurements correspond to concepts with the lowest popularity/trend scores*. In particular, concepts like *Mona Lisa*, *climate change* and *Battle of Stalingrad* have very low $t(C)$ scores (0.13, 0.33, 0.09) but very high concept drift (1.56, 1.73, 1.74). Contrarily, concepts with the highest $t(C)$ scores, such as *Donald Trump* (8.99), *Pokemon* (2.29) and *iPhone* (6.75)), have increasing values of concept drift (1.53, 1.16, 1.36) but never reach that of the non-trendy concepts. Less popular, but very trendy concepts such as *Donald Trump* change their relevance when observing $p(C)$, but the tendency to score less concept drift prevails.

These two unexpected patterns could be explained by the *experts vs crowds* hypothesis. Under this hypothesis, most significant edits in Wikipedia in a concept C (which derive in high drift scores) are poorly explained by querying trends over C, but much related to a tiny amount of Wikipedia curators (the “experts”) taking care of domain-expert content (i.e. *Mona Lisa*, *Battle of Stalingrad*). So, experts would be responsible of concept drift in less trendy topics. However, the “crowds” seem to be able to influence concept drift approximately linearly (*Pokemon*, *iPhone*, *Donald Trump*) beyond a certain trendiness threshold. This would explain high-quantity/low-quality edits in Wikipedia derived from controversy and popularity, and relate to the popularity required to score some increasing concept drift by non-experts. Despite this, highest trend values do not seem to involve deep intensional changes in concepts, which only occur in expert curated, low-trendiness concepts.

5 CONCLUSION AND FUTURE WORK

In this paper, we study the influence of trending Web queries over the fundamental properties of collaborative Web knowledge bases. In the period of 2014 January-2016 April and a small sample of concepts with variable popularity, we find patterns that fit the possible explanation of two conflicting trends (“experts vs. crowds”) with competing influence on intensional concept drift. We plan to add scalability to our framework in order to confirm the above findings, and to investigate automatic mapping methods between concepts and their corresponding search-term queries.

REFERENCES

[1] Alexander Mäedche, Boris Motik, and Ljiljana Stojanovic. 2003. Managing multiple and distributed ontologies in the Semantic Web. *The VLDB Journal — The International Journal on Very Large Data Bases* 12, 4 (2003), 286–300.

[2] J. Allan, J. Carbonell, G. Doddington, J. Yamron, and Y. Yang. 1998. Topic detection and tracking pilot study: Final report. In *Proceedings of the DARPA Broadcast News Transcription and Understanding Workshop.*

[3] Tim Berners-Lee, James Hendler, and Ora Lassila. 2001. The Semantic Web. *Scientific American* 284, 5 (2001), 34–43.

[4] S. Darányi and P. Wittek. 2013. Demonstrating Conceptual Dynamics in an Evolving Text Collection. *Journal of the American Society for Information Science and Technology* 64, 12 (2013), 2564–2572. DOI:http://dx.doi.org/10.1002/asi.22940

[5] G.P.C. Fung, J.X. Yu, P.S. Yu, and H. Lu. 2005. Parameter free bursty events detection in text streams. In *Proceedings of VLDB-05, 31st International Conference on Very Large Data Bases.* Trondheim, Norway, 181–192.

[6] E. Gabrilovich, S. Dumais, and E. Horvitz. 2004. Newsjunkie: providing personalized newsfeeds via analysis of information novelty. In *Proceedings of WWW-04, 13th Int. Conf. on the World Wide Web.* New York City, NY, USA, 482–490.

[7] R. S. Gonçalves, B. Parsia, and U. Sattler. 2011. Analysing Multiple Versions of an Ontology: A Study of the NCI Thesaurus. In *Proceedings of the 24th Int. Workshop on Description Logics (DL 2011)*, Vol. 745. CEUR Workshop Proceedings.

[8] Z. Harris. 1970. Distributional structure. In *Papers in structural and transformational Linguistics*, Z. Harris (Ed.). Humanities Press, NY, USA, 775–794.

[9] Tom Heath and Christian Bizer. 2011. *Linked Data: Evolving the Web into a Global Data Space* (1st ed.). Morgan and Claypool. 1–136 pages.

[10] Bryan Jurish. 2016. Diachronic Collocations and Genre: a case for DiaCollo?. In *Diachronic Corpora, Genre, and Language Change*, Richard Jason Whitt (Ed.). 22–24. http://kaskade.dwds.de/~jurish/pubs/jurish2016genre.pdf

[11] J. Kleinberg. 2003. Bursty and hierarchical structure in streams. *Data Mining and Knowledge Discovery* 7, 4 (2003), 373–397.

[12] J. Kleinberg. 2006. Temporal dynamics of on-line information streams. *Data Stream Management: Processing High-Speed Data Streams* (2006).

[13] P. De Leenheer and T. Mens. 2008. Ontology Evolution: State of the Art and Future Directions. In *Ontology Management for the Semantic Web, Semantic Web Services, and Business Applications.* Springer.

[14] J. Lehmann, R. Isele, M. Jakob, A. Jentzsch, D. Kontokostas, P. N. Mendes, S. Hellmann, M. Morsey, P. van Kleef, S. Auer, and C. Bizer. 2014. DBpedia - A Large-scale, Multilingual Knowledge Base Extracted from Wikipedia. *Semantic Web – Interoperability, Usability, Applicability* (2014). http://www.semantic-web-journal.net/system/files/swj558.pdf.

[15] Albert Meroño-Peñuela, Christophe Guéret, Ashkan Ashkpour, and Stefan Schlobach. 2015. CEDAR: The Dutch Historical Censuses as Linked Open Data. *Semantic Web – Interoperability, Usability, Applicability* (2015). In press.

[16] R. Papka. 1999. *On-line new event detection, clustering, and tracking.* Ph.D. Dissertation. University of Massachusetts Amherst.

[17] Catia Pesquita and Francisco M. Couto. 2012. Predicting the Extension of Biomedical Ontologies. *PLoS Computational Biology* 8, 9 (2012), e1002630.

[18] Michael Petrilli. 2008. Wikipedia or Wickedpedia? http://educationnext.org/wikipedia-or-wickedpedia/, *Education Next* 8, 2 (2008).

[19] T. G. Stavropoulos, S. Andreadis, E. Kontopoulos, M. Riga, P. Mitzias, and I. Kompatsiaris. 2017. The SemaDrift Protégé Plugin to Measure Semantic Drift in Ontologies: Lessons Learned. In *Knowledge Engineering and Knowledge Management (EKAW 2016)*, Vol. 10180. Springer, Cham, 29–39.

[20] T. G. Stavropoulos, S. Andreadis, M. Riga, E. Kontopoulos, P. Mitzias, and I. Kompatsiaris. 2016. A Framework for Measuring Semantic Drift in Ontologies. In *Proceedings of SuCCESS-16, 1st Int. Workshop on Semantic Change & Evolving Semantics, co-located with the 12th European Conference on Semantics Systems (SEMANTiCS-16).*

[21] T. G. Stavropoulos, E. Kontopoulos, A. Meroño Peñuela, S. Tachos, S. Andreadis, and I. Kompatsiaris. 2017. Cross-domain Semantic Drift Measurement in Ontologies Using the SemaDrift Tool and Metrics. In *3rd Workshop on Managing the Evolution and Preservation of the Data Web (MEPDaW 2017).*

[22] Ljiljana Stojanovic. 2004. *Methods and Tools for Ontology Evolution.* Ph.D. Dissertation. University of Karlsruhe.

[23] Ljiljana Stojanovic and Boris Motik. 2002. Ontology Evolution within Ontology Editors. In *Evaluation of Ontology-based Tools Workshop, 13th Int. Conf. on Knowledge Engineering and Knowledge Management (EKAW 2002)*, Vol. 62. CEUR-WS.

[24] R. Swan and J. Allan. 1999. Extracting significant time varying features from text. In *Proceedings of CIKM-99, 8th International Conference on Information and Knowledge Management.* Kansas City, MO, USA, 38–45.

[25] R. Swan and D. Jensen. 2000. Timemines: Constructing timelines with statistical models of word usage. In *Proceedings of KDD-2000 Workshop on Text Mining.* Boston, MA, USA, 73–80.

[26] Ilaria Tiddi. 2016. *Explaining Data Patterns using Knowledge from the Web of Data.* Ph.D. Dissertation. Knowledge Media Institute, The Open University.

[27] R. Verborgh, M. van der Sande, O. Hartig, J. Van Herwegen, L. De Vocht, B. De Meester, G. Haesendonck, and P. Colpaert. 2016. Triple Pattern Fragments: a Low-cost Knowledge Graph Interface for the Web. *Journal of Web Semantics* 37–38 (2016), 184–206. DOI:http://dx.doi.org/doi:10.1016/j.websem.2016.03.003

[28] S. Wang, S. Schlobach, and M. C. A. Klein. 2010. What Is Concept Drift and How to Measure It?. In *Knowledge Engineering and Management by the Masses - 17th Int. Conf., EKAW 2010. Proceedings.* LNCS 6317, Springer, 241–256.