

EventMedia Live: Reconciling Events Descriptions in the Web of Data

Houda Khrouf and Raphaël Troncy

EURECOM, Sophia Antipolis, France

<houda.khrouf@eurecom.fr>

<raphael.troncy@eurecom.fr>

1 Introduction

Many online services provide functionalities for sharing one’s participation and captured media at real-world events. In a previous work [3], we constructed the EventMedia dataset aggregating heterogeneous sources of information and producing linked data. In this work, we carry out an event-oriented data reconciliation experiment in the web of data, and we propose two adapted extensions to a semi-automatic alignment tool named SILK [2].

2 Events Reconciliation

In this work, we aim at creating high quality owl:sameAs links between similar events which share an overlap in term of three properties: title, location and date. Hereafter, we propose two similarity extensions that better comply with event properties as it will be confirmed by some experimental results.

Temporal Inclusion metric. Intuitively, we consider that two events are similar if they share among others the same time or temporal interval. We introduce a new temporal inclusion metric where we define a parameter θ as the number of hours that can be tolerated between two dates. Given two events (e_1, e_2) which have respectively the couple start date and end date (d_1, d'_1) and (d_2, d'_2) where $d_1, d_2 \neq 0$ and d'_1, d'_2 can be null, the temporal inclusion (s) metric is defined by:

$$s(e_1, e_2) = \begin{cases} 1 & \text{if } |d_1 - d_2| \leq \theta \text{ where } (d'_1, d'_2) = 0 \\ 1 & \text{if } d_1 \pm \theta \in [d_2, d'_2] \text{ where } d'_1 = 0 \text{ (idem for } d_2) \\ 1 & \text{if } \min(d'_1, d'_2) - \max(d_1, d_2) \geq 0 \text{ where } (d'_1, d'_2) \neq 0 \\ 0 & \text{otherwise.} \end{cases} \quad (1)$$

Token-Wise string similarity. We studied the performance of string similarity functions over the EventMedia dataset that consists of user-generated content featuring typos and noisy data. We introduce a novel metric called Token-Wise combining the character-based and token-based string similarity metrics. The strings s and t are firstly split into a set of tokens $s_1 \dots s_k$ and $t = t_1 \dots t_p$. Given $s'(s_i, t_j)$ the score of the based-character similarity between two tokens, ws_i and wt_j the respectively weights of s_i and t_j , N the number of matched tokens

(filtered triples), M the number of unmatched tokens where we set $s' = 0$, the token-wise (tw) metric is defined by:

$$tw(s, t) = \frac{\sum_{i=1}^N s'(s_i, t_j) \times ws_i \times wt_j}{\sum_{i=1}^{N+M} (1 - s'(s_i, t_j)) \times (ws_i^2 + wt_j^2) + \sum_{i=1}^N s'(s_i, t_j) \times ws_i \times wt_j} \quad (2)$$

Experimental Results. The first experiment was applied on agents’ names to compare the token-wise and Jaro [1] distances based on a ground truth of 150 matched instances between Last.fm and MusicBrainz. The table 1 shows the recall and precision for different thresholds. In the second experiment, we evaluate the event alignment approach based on a ground truth containing 68 Eventful events compared with 104 Upcoming events, and 583 Last.fm events compared with 533 Upcoming events. Table 2 shows the recall and precision when the parameter θ of temporal inclusion is equal to 0 and 24 hours.

μ	Jaro		Token-Wise	
	Recall(%)	Precision (%)	Recall (%)	Precision(%)
0.95	24	100	60	100
0.9	49	98	82	99
0.8	87	93	96	98
0.7	100	45	100	77

Table 1. Precision and Recall for Jaro and Token-Wise similarities

$\mu >$	Eventful-Upcoming				LastFm-Upcoming			
	$\theta = 0$		$\theta = 24 H$		$\theta = 0$		$\theta = 24 H$	
	Recall	Precision	Recall	Precision	Recall	Precision	Recall	Precision
0.8	43	100	71	100	41	100	87	100
0.75	48	100	74	93	43	100	90	100
0.74	74	25	79	26	74	81	94	85
0.6	100	24	100	26	100	75	100	75

Table 2. Precision and Recall for events alignment

3 Conclusion

We proposed a powerful string similarity metric to cope with noisy titles, and a temporal inclusion similarity metric detecting a temporal overlap. The evaluation shows good results consolidating the efficiency of these extensions for SILK.

References

1. Matthew A. Jaro. Advances in Record-Linkage Methodology as Applied to Matching the 1985 Census of Tampa, Florida. *Journal of the American Statistical Association*, 84(406):414–420, 1989.
2. A. Jentzsch, R. Isele, and C. Bizer. Silk - Generating RDF Links while publishing or consuming Linked Data. In *9th International Semantic Web Conference (ISWC’10)*, Shanghai, China, 2010.
3. R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS’10)*, Graz, Austria, 2010.