Temporal Reconciliation Based on Entity Information

Paul Martin, Marc Spaniol, Antoine Doucet

▶ **To cite this version:**

Paul Martin, Marc Spaniol, Antoine Doucet. Temporal Reconciliation Based on Entity Information. International Symposium on Web AlGorithms, Jun 2015, Deauville, France. hal-01171336

HAL Id: hal-01171336
https://hal.science/hal-01171336

Submitted on 8 Jul 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Temporal Reconciliation Based on Entity Information

Paul MARTIN
Normandy University
paul.martin@unicaen.fr

Marc SPANIOL
Normandy University
marc.spaniol@unicaen.fr

Antoine DOUCET
University of La Rochelle
antoine.doucet@univ-lr.fr

Abstract

Temporal classification of Web contents requires a "notion" about them. This is particularly relevant when contents contain several dates and a human "interpretation" is required in order to chose the appropriate time point. The dating challenge becomes even more complex, when images have to be dated based on the content describing them. In this paper, we present a novel time-stamping approach based on semantics derived from the document. To this end, we will first introduce our experimental dataset and then explain our temporal reconciliation pipeline. In particular, we will explain the process of temporal reconciliation by incorporating information derived from named entities.

I. Introduction

In the era of the Web 2.0 and its social media, a "new" kind of user has emerged: the prosumer (producer and consumer at the same time). These users produce digital contents at a tremendous speed across all social media platforms. Commonly, these contents also contain images, which come from various sources, such as modern digital cameras but also result from digitalization by scanners. As such, these images cover different periods of time and carry interesting temporal information about places, people, events, facts, *etc.*

A problem inherent to social media contents and - in particular - the images they contain is proper dating. Although novel cameras and scanners enable users to automatically date the contents, in reality this information is often faulty since devices are not properly set-up. Particularly in digitalization the date captured is often *the digitalization date rather than the real taken date.* This leads to contents that carry inconsistent temporal information and thus cannot be properly exploited.

II. Conceptual Approach

Our hypothesis is that the *real* temporal information of a photo is carried inside both the image itself and - if existing - the surrounding text. However, "time-stamping" an image based on its surrounding text is a non-trivial task: the date might not be explicitly mentioned, it may describe a different aspect or there may actually be several dates from which to choose. In such cases, a human is commonly able to "interpret" the content and identify the proper time point. To this end, the human exploits the semantics of the content and applies the knowledge about it for temporal alignment.

Table 1: *Dataset Statistics*

Collection and Annotations	
towns (all areas)	50
years covered (from-to)	190 (1820-2010)
annotations made	55737
positive street annotations	15944
images annotated as street	6947
Date obtained from	
inside the page	4071 (25.53%)
image file name	1629 (10.22%)
image alt	76 (0.48%)
page title	69 (0.43%)
image title	46 (0.29%)
other:not found, errors *etc.*	10053 (63.05%)

With the emergence of knowledge bases such as DBpedia [1], Freebase [2], or YAGO [9] factual knowledge about named entities has become machine-readable. We therefore "mimic" the human behavior by exploiting information derived from the named entities for temporal reconciliation.

III. Dataset

We collected a dataset of Web pages containing images retrieved by google images according to a particular set of "temporal queries". To build these queries we have selected a set of 50 towns that were elected in a MasterCard study of the *top 10 Destination Cities by International Overnight Visitors* inside 5 regions of the globe. We followed the idea from Dias *et al.* [3] to build the queries by adding to each town name a year in the time-range from 1820 to 2010: *Query* = "*a town name*" + "*a year*" .

We employed 3 annotators over a random subset of documents. For images they identified as street images, they were further asked to identify the proper year. Table 1 gives an overview on the extracted information.

IV. Temporal Reconciliation

We now describe our reconciliation pipeline.

a) **Clean pages and extract image position** We have cleaned the pages from scripts, html tags, styles... with a series of regular expressions and converted them into UTF-8 encoding. Simultaneously we have detected and stored the position of the image in the original page and in the clean page.

b) **Extract temporal information and positions** In order to extract temporal information such as dates, decades or periods of time from text we use regular expressions of 3 forms respectively the *YYYY*, *YYY0s* and *YYYY-YYYY* forms.

c) **Extract named entities and explore temporal facts** Extraction of semantic temporal information is based on AIDA [6], which disambiguates named entities onto the YAGO2 knowledge base [5]. These named entities enable us to extract semantics about the content. In particular, we incorporate relations that carry temporal information inside the *yagoLiteralFacts* (such as "*wasDestroyedOnDate*"), which we use for temporal reconciliation.

V. Initial Results

Based on the data set described in Section III we conducted experiments on the dating of images. As a "baseline" we implemented an approach that dates the image based on the closest time mentioned in the surrounding text. With an accuracy of about 61% this baseline implementation already gives a fairly good quality. When incorporating our temporal reconciliation method based on the temporal knowledge derived from the contained named entities as described in Section IV, initial experiments show that temporal reconciliation significantly improves precision.

VI. Related Research

Early methods in temporal classification of images where manual methods based on the physical photographic medium that have evolved through time. Kodak proposed the first automatic method using both sides of the image by detecting watermarks or paper references over or back-printed on the images. The first automatic method using visual features has been proposed recently by Palermo *et al.* [8]. Following this approach other temporal color evolution feature based methods were proposed by [4]. Martin *et al.* [7] proposed an ordinal classification framework which addresses the problem in a more adapted manner focusing on the ordinal nature of this problem.

VII. Conclusion and Outlook

In this paper we presented a novel timestamping approach that "mimics" human behavior by exploiting information derived from the named entities. As we have reported in Section V our initial results are quite promising and show significant improvement. However, this temporal reconciliation comes at a price: while we gain in precision, we loose in recall. Future research now aims at increasing both precision and recall.

Acknowledgements

We are grateful to Frédéric Jurie for his productive comments and discussions.

References

[1] Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, and Zachary Ives. Dbpedia: A nucleus for a web of open data. In *In 6th Intl Semantic Web Conference, Busan, Korea*, pages 11–15. Springer, 2007.

[2] Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. Freebase: A collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, pages 1247–1250, New York, NY, USA, 2008. ACM.

[3] Gaël Dias, José G. Moreno, Adam Jatowt, and Ricardo Campos. Temporal web image retrieval. In *19th International Conference on String Processing and Information Retrieval*, SPIRE'12, pages 199–204, 2012.

[4] Basura Fernando, Damien Muselet, Rahat Khan, Tinne Tuytelaars, and KU PSI-VISICS. Color features for dating historical color images. In *ICIP*, 2014.

[5] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, and Gerhard Weikum. YAGO2: a spatially and temporally enhanced knowledge base from wikipedia. Research Report MPI-I-2010-5-007, Max-Planck-Institut für Informatik, Stuhlsatzenhausweg 85, 66123 Saarbrücken, Germany, November 2010.

[6] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, pages 782–792, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[7] Paul Martin, Antoine Doucet, and Frédéric Jurie. Dating color images with ordinal classification. In *Proceedings of International Conference on Multimedia Retrieval*, page 447. ACM, 2014.

[8] Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *ECCV*, pages 499–512. Springer, 2012.

[9] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proc. of WWW*, pages 697–706, 2007.