Semantic Annotation to Support Description of the Art Market

Dominik Filipiak, Krzysztof Węcel, Agata Filipowska
Department of Information Systems, Poznań University of Economics
al. Niepodległości 10
61-875 Poznań, Poland
dominik.filipiak,krzysztof.wecel,agata.filipowska@kie.ue.poznan.pl

ABSTRACT

The estimation of prices on the art market has been investigated as a research topic for many years, but only recently new approaches to this problem have been applied. One of these approaches concerns extending data on a work of art with data from the Internet to improve the quality of assessment. This, however, creates a lot of challenges mostly regarding the information extraction. Semantic annotation and enrichment of the crawled data enable additional reasoning and introduce new features into existing methods, resulting in a better estimation of indices for the art market.

The problem tackled by this paper is as follows: what kind of semantic enrichment on documents collected from the Internet can be introduced to extend the data on the artwork and influence the efficiency and quality of indices calculated for artworks.

Categories and Subject Descriptors

I.2 [**Artificial Intelligence**]: Natural Language Processing; J.5 [**Arts and Humanities**]

General Terms

Econometrics

Keywords

Art Market Analysis, Linked Data, Semantic Tagging, Econometrics

1. INTRODUCTION

1.1 Art Market

With all their *invaluable* qualities, artworks are often treated as a type of an asset, just like stocks or bonds. The idea of considering art as a form of an alternative investment can be perceived as a controversial one. Nonetheless, this approach to artworks is recently gaining more and more attention[1]. As a consequence, numerous studies have been carried out to explore the topic.

Surprisingly, artworks (such as paintings) traded in auction houses can be described with a decent number of variables, regardless of one's art history knowledge. The name of an author, medium used, size of a painting, initial and hammer price, to name a few. This data is often provided by institutions taking care of the sale. Sometimes, among other features, a long text description is associated with the painting or artist. This is a valuable source of information, but as it is provided in a natural language, it has to be processed before machine-based data processing is possible.

1.2 Problem Statement

Since auction houses have started to frequently publish sales results on the Internet, the perception of the art market has changed. Numerous services have started to collect data and even prepare market reports. Minimisation of information asymmetry was not the only consequence of this step. With a sufficient amount of high-quality data, research carried out on the art market finally can be conducted in a data science manner. An employment of a hedonic regression using semantically enriched information is presented in this paper as a base for building art market indices. The concept of using a regression in the art market analysis has been intensively studied, but, to the best of our knowledge, the use of semantic data enrichment constitutes a contribution to this field.

The problem tackled by this paper was defined as follows: what kind of semantic enrichment on data published on the Internet, e.g. by auction houses, can be introduced to extend the data on the artwork and influence the efficiency and quality of the calculation of artworks' indices.

2. SOLUTION

2.1 Architecture of the Solution

In order to tackle the presented problem, a solution offering a four-step processing pipeline has been designed. These steps concern data collection, data refinement, information extraction and data enrichment. The solution may be per-

[1]An interesting case is provided by the National Bank of Hungary, which started to invest in art works `http://blogs.wsj.com/emergingeurope/2014/03/31/hungary-central-bank-to-buy-art/`

ceived as a sort of framework, which is a base for future research.

A reasonable number of observations regarding sales in auction houses must be collected to facilitate building of an effective prediction model. Therefore, the first step considers the data collection. Numerous services provide historical sales information. *Artprice*[2] and *Artnet*[3] are the most prominent examples of data providers. However, these sites are often subscription-based and do not provide data in a parse-friendly format, not to mention various legal issues. As a consequence, the data collection must be performed by dedicated crawlers, operating on pages of numerous auction houses.

Data refinement and cleansing is indispensable in order to obtain robust results. For example, due to a human error, some observations have misspelled information about artists. This issue may be resolved by applying various fuzzy string matching algorithms. According to the so-called *garbage in, garbage out* principle, this step is crucial to assure the quality of the experiment's results.

The third step considers information extraction from the collected documents. Auction lots are often described with an unstructured text which contains useful information. For instance, the presence of a signature on a painting or a number in an edition in lithographies may carry important information influencing the hammer prices. Due to the complexity of this process, possible approaches are described in detail in section 2.2.

The data enrichment, the final step, makes use of annotated entities in order to provide more complementary information. Although minimising information asymmetry on the art market is an obvious goal behind this approach, there are various applications of enriched data. These possibilities are covered in section 2.3.

2.2 Annotation – From Text to Triples

Ontologies used for annotation make it easier for people and machines to understand the text. Document retrieval can be significantly improved when additional relations from ontology are leveraged. For example, we can ask for documents containing information about impressionists and we actually do not have to know the names of individual artists. Also, any document containing the phrase "oil painting" will be classified as a "document about art media".

DBpedia Spotlight[4] is a tool for automatically annotating mentions of DBpedia resources in text. It classifies entities according to the DBpedia ontology. Two modes are available. In the first one (candidates), it spots the potential mentions (either statistically or based on gazetteer) and retrieves the candidate DBpedia resources bound to Wikipedia. In the second mode (annotate), it additionally disambiguates candidates and links the mentions to the best one. One of the strong points of the DBpedia Spotlight is the richness of language resources that can be used for indexing by the underlying engine. Depending on the language sometimes additional processing is required to improve recall of spotting. For example, for the Polish language an external morphology analyser is necessary to normalise various inflectional forms.

Several solutions base also on considering many ontologies at once. NERD – Named Entity Recognition and Disambiguation[5] proposes unified numerous named entity extractors using the NERD ontology, which provides a set of axioms aligning various underlying taxonomies. Mappings are established manually. According to the documentation, several extractors are supported, including: DBpedia Spotlight, OpenCalais and Zemanta.

A similar "meta-approach" is taken by Apache Stanbol, a general framework for semantic enhancement of unstructured text. DBpedia Spotlight can work as an EnhancementEngine for Stanbol. Stanbol also links to several other external services via enhancement modules, for example: Named Entity Linking Engine (suggests links to linked data sources), FST Linking Engine (links Entities indexed in a Solr index), Geonames Enhancement Engine (links to geonames.org, with hierarchical links for locations), OpenCalais (both NER and Entity Linking), Zemanta Enhancement Engine (both NLP and Entity Linking).

One of the best solutions regarding disambiguation is the Dandelion[6] service offered by Spaziodati. The integration is much deeper than in the case of NERD where only ontology was aligned. It builds truly own knowledge graph which allows for much better ranking and thus more precise disambiguation of the mentions. Figure 1 presents a sample annotation of an artwork with the Dandelion API.

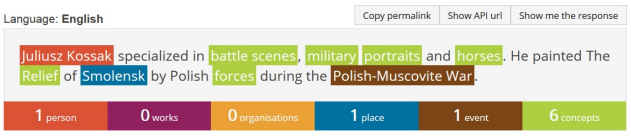

Figure 1: Named entity recognition by Dandelion API

2.3 Data Enrichment

Sometimes it is hard to distinguish data extraction from enrichment; very often these phases are combined. Having identified the entity in text, additional data can be retrieved. Semantic enrichment sometimes covers phases three and four of the proposed approach, being information extraction and data enrichment stemming from semantic annotation. In the context of data mining links to external information results in additional attributes, they can improve the quality of the predictive model [1].

There is a big number of open data sources that may enrich data on artworks currently available on the Web. The domain of fine art requires more thoughtful selection as not all datasets contain relevant information. The obvious source is DBpedia[7]. The recent version (2014) provides informa-

[2] http://www.artprice.com
[3] https://www.artnet.com/price-database/
[4] http://spotlight.dbpedia.org
[5] http://nerd.eurecom.fr/
[6] https://dandelion.eu/
[7] http://dbpedia.org

tion about more than 1,445,000 people and 411,000 creative works in its English edition only. Other language chapters can potentially provide additional data, particularly, when the DBpedia language matches the nationality of the artist.

An intensive effort has been observed to align artists' descriptions in Wikipedia with external authoritative sources, when possible. Therefore, we can expect more precise and more complete information about at least some of the artists. At the bottom of the Wikipedia article for some people one can find "Authority control" with links to external sources, e.g. VIAF (Virtual International Authority File), ISNI (International Standard Name Identifier), ULAN (Union List of Artist Names). The number of available references depends on the popularity of the artist. For example, there are 11 entries for the famous Polish painter Wojciech Kossak and only four for his lesser-known son (see Figure 2).

Figure 2: Authority control record for Wojciech Kossak

VIAF is a joint project of several national libraries. It is apparently the biggest dataset, with information on about over 35 million names[8]. The provenance information is kept for each piece of data, which is useful as some discrepancies still exists. For example, Figure 3 presents different dates of birth of Wojciech Kossak.

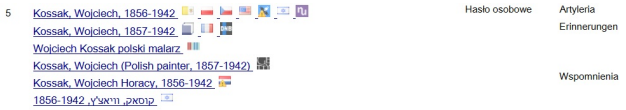

Figure 3: VIAF records for Wojciech Kossak

ISNI is an ISO 27729 global standard number for identifying contributors to creative works. It currently (2015) holds the information about more than 8 million individuals. ULAN[9] is particularly interesting as it focuses on artists' names, holding information on about 120,000 of them. The basic information includes given names (in multiple languages), pseudonyms and variants spelling, i.e. various surface forms (almost 300,000). Such information is crucial for finding mentions of an artist in the text. The search interface allows to find all artists with a given name (Figure 4) and within individual page relations between artists are also provided (Figure 5).

Another dataset offered by Getty is also relevant to our research – The Art & Architecture Thesaurus (AAT)[10]. It contains terms useful in the description of art techniques (see

[8]VIAF Annual Report 2014, `http://www.oclc.org/content/dam/oclc/viaf/OCLC-2014-VIAF-Annual-Report-to-VIAF-Council.pdf`

[9]`http://www.getty.edu/research/tools/vocabularies/ulan/index.html`

[10]`http://www.getty.edu/research/tools/vocabularies/aat/index.html`

Figure 4: ULAN search results for 'Kossak'

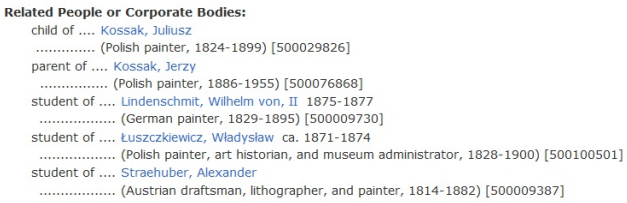

Figure 5: Information about Wojciech Kossak in ULAN

Figure 6). A set of Getty's vocabularies is available as linked open data (LOD)[11] making it a perfect fit for supplementing annotations in an ontological form.

As we keep data in Open Refine, it would be convenient to use one of the extensions for named entity recognition. That would allow us to extend our data about a certain artwork with additional attributes, thus leading to better predictive models. In such context RDF-extension (developed by DERI Galway) with such functionalities as reconciliating against SPARQL endpoints or RDF dumps and exporting to RDF might be used. DBpedia-extension (by Zemanta) added the possibility to extend reconciled data with data from DBpedia and to extract entities from full text descriptions via Zemanta API. Regarding integration, one of the most comprehensive solutions is LODGrefine, developed within thr LOD2 project[12]. Unfortunately, it is targeted at the English language, and we need to adapt it for Polish. It also does not contain domain-specific ontologies like for example ULAN.

To conclude, the way various tools conduct analysis is very similar. In fact, only two aspects make these solutions different: the underlying dictionary and the ability to disam-

[11]`http://www.getty.edu/research/tools/vocabularies/lod/`

[12]`http://lod2.eu`

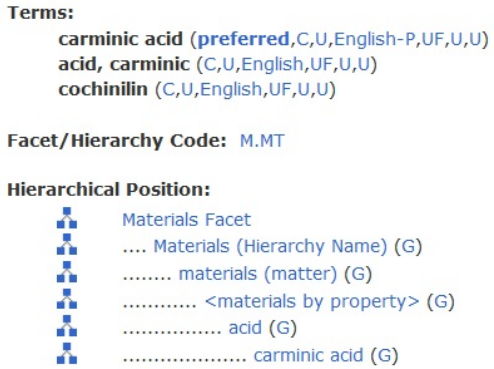

Figure 6: AAT classification about cochinilin pigment

biguate entities. None of the solutions offer a direct support for the Polish language. These aspects open a space for our improvements. Our solution will base on the DBpedia Spotlight with an index built for the Polish language resources from the Polish DBpedia supplemented with domain-specific ontologies like ULAN.

3. MARKET INDICES PREDICTION

The semantically annotated data used while describing art works may improve the process of creation of indices for the art market. The art market indices are build for outlining general trends and measure its volatility and overall value. Comparison of artworks with more traditional forms of assets (like bonds) or searching for a correlation between various economic factors and behaviour of the market complement the rationale behind constructing indices [2].

Currently, two ways to develop art market indices are the most popular: based on repeat-sales and hedonic regression. The first method takes into account all items sold at least twice and calculates indices based on the proportion of the first and the second sale prices. Probably the most notable example of this approach is the Mei&Moses Art Index [4]. Its weakness relates to the fact that artworks are considered as a long-term form of investment, what results in a relatively small amount of data to base on. Therefore, many researches have employed hedonic regression [3]. It is a form of linear regression, which takes into account various features of artworks and their year of sale separately compared to the auction lot hammer price in this case. The Ordinal Least Squared method is used to estimate coefficients. A simple example of this linearised model is presented in equation 1.

$$\ln P_{it} = \alpha + \sum_{j=1}^{z} \beta_j X_{ij} + \sum_{t=0}^{\tau} \gamma_t D_{it} + \varepsilon_{it}, \qquad (1)$$

where $\ln P_{it}$ represents the natural logarithm of a price of a given painting $i \in \{1, 2, ..., N\}$ at time $t \in \{1, 2, ..., \tau\}$; α, β and γ are regression coefficients for estimated characteristics. X_{ij} represents hedonic variables included in the model, whereas D_{it} stands for time dummy variables.

Considered hedonic variables are, for example, the artist's name, the painting's size, year of creation and other related features describing a given painting. An indirect information, such as a time of death of the artists, may also be included. In some cases information can be missing, therefore this method is considered to be prone to the selection bias. Having a wide range of relevant data is one of the most important steps in the index calculation process. Therefore, this is the place where the approach discussed in the paper can be used for yielding more accurate indices. More complete data with extracted variables (such as the mentioned presence of a signature or edition in the case of lithographies) allows to build more sophisticated representation of a painting. Used in the equation (1), it results in more accurate coefficients representing various sales periods (γ_t). These coefficients are actually employed to construct indices:

$$Index_{t+1} = \frac{e^{\gamma_t}}{e^{\gamma_{t+1}}} \qquad (2)$$

which can be used to measure and visualise overall art market performance through different periods.

4. CONCLUSIONS

Nowadays, we deal with an increasing popularity of investment in artworks. This imposes the need for employing various methods for estimation of prices of these artworks. Therefore, researchers and practitioners work on methods enabling market description and price estimation. This relates also to indices developed for the art market that were addressed in the paper.

The paper presented the approach of how the semantic processing may enrich data available for the current methods of estimation of indices for the art market. It discussed the data sources as well as proposed the semantically-based document processing pipeline. Currently, this approach is being implemented and the first results seem promising.

Having annotated descriptions of artists and artworks, it is possible to conduct further research. A complementary, detailed and semantically enriched *catalogue raisonné* obtained in the previously mentioned process could be a valuable source of information for performing art market analysis itself. In addition, well-structured data may pave the way towards usage of methods from a graph theory, topic labelling or even employment of machine learning.

5. REFERENCES

[1] C. d'Amato, P. Berka, V. Svátek, and K. Węcel, editors. *Proc. of the International Workshop on Data Mining on Linked Data collocated with (ECMLPKDD 2013), Prague, Czech Republic, Sep. 23, 2013*, volume 1082 of *CEUR Workshop Proceedings*. CEUR-WS.org, 2013.

[2] V. Ginsburgh, J. Mei, and M. Moses. The Computation of Prices Indices. In *Handbook of the Economics of Art and Culture*, volume 1, pages 947–979. Elsevier, 2006.

[3] R. Kräussl and N. van Eisland. Constructing the True Art Market Index - A Novel 2-Step Hedonic Approach and its Application to the German Art Market. 2008.

[4] J. Mei and M. Moses. Art as an Investment and the Underperformance of Masterpieces. *NYU Finance Working Paper*, (FIN-01-012):1–23, 2001.