AESTHETICS: Analytics with Strings, Things, and Cats

Johannes Hoffart
Max Planck Institute for Informatics
jhoffart@mpi-inf.mpg.de

Dragan Milchevski
Max Planck Institute for Informatics
dmilchev@mpi-inf.mpg.de

Gerhard Weikum
Max Planck Institute for Informatics
weikum@mpi-inf.mpg.de

ABSTRACT

This paper describes an advanced news analytics and exploration system that allows users to visualize trends of entities like politicians, countries, and organizations in continuously updated news articles. Our system improves state-of-the-art text analytics by linking ambiguous names in news articles to entities in knowledge bases like Freebase, DBpedia or YAGO. This step enables indexing entities and interpreting the contents in terms of entities. This way, the analysis of trends and co-occurrences of entities gains accuracy, and by leveraging the taxonomic type hierarchy of knowledge bases, also in expressiveness and usability. In particular, we can analyze not only individual entities, but also categories of entities and their combinations, including co-occurrences with informative text phrases. Our Web-based system demonstrates the power of this approach by insightful anecdotic analysis of recent events in the news.

1. MOTIVATION AND INTRODUCTION

Proper analysis and understanding of large amounts of texts has become a major necessity, as the amount of natural-language contents keeps growing, especially in the context of social media and daily news, but also regarding scholarly publications and digitized books. A notable project in this line is the Culturomics [9] project, which supports an aggregated view of trends in the recent human history captured by the Google books corpus. Here, interesting conclusions are drawn about linguistic changes or cultural phenomena using string-level keywords, by comparing frequencies over time periods and across languages.

The system presented in this paper, called AESTHETICS (short for Analysis and Exploration with Strings, Things, and Categories), goes beyond mere string-based analysis, by supporting the analysis and exploration of entities ("things") and categories ("cats"). This semantic level is provided by linking text phrases to knowledge bases like Google's Knowledge Graph (`freebase.com`) or `yago-knowledge.org`. Knowledge bases contain a large number of persons, places, organizations, etc. providing a repository of unique entities in canonicalized form, with assignment to fine-grained semantic classes (categories). The AESTHETICS system automatically discovers and disambiguates these entities in news texts, linking the unstructured to the structured world. Instead of specifying words or phrases as the target of mining trends and patterns, we can now see and analyze entities directly. AESTHETICS supports the analysis of textual surface phrases ("strings") as well, but its full power comes from combining these with proper entities and categories.

To illustrate why this is a major step with strong benefits, consider the task of visualizing trends around the recent Ukrainian crisis, which originated from the Maidan, the square in Kiev where thousands of Ukrainians protested in early 2014. A search for "Maidan" quickly reveals that the name is highly ambiguous, as it means "square" not only in Ukrainian, but also in Hindi and Arabic. Thus, simply counting the string "Maidan" will result in a large number of false positives, leading to an imprecise analysis, as shown in Figure 1. By specifying the canonicalized entity `Maidan Nezalezhnosti`, not only do we get rid of spurious mentions of other Maidans, but also find articles where the square is mentioned only by its English name "Independence Square". Thus, entity-level analytics, as supported by AESTHETICS, is the only way to get accurate numbers.

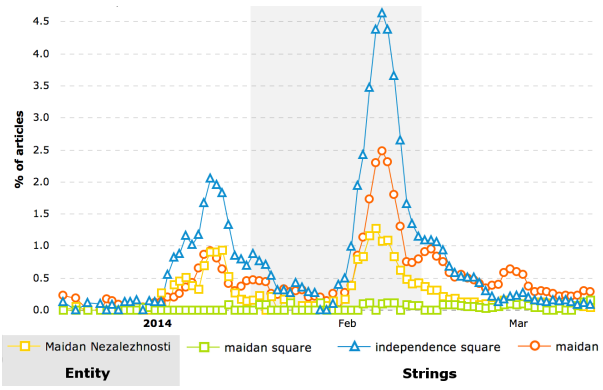

Figure 1: Accurate analytics for Maidan Square

Additionally, as we now have the full potential of a structured knowledge base in the background, further opportunities are opened up. In all semantic knowledge bases, entities are organized in a category hierarchy, e. g. `Greenpeace` is an `environmental organization`, which in turn is a subclass

CIKM'14, Nov 03–07, 2014, Shanghai, China.
ACM 978-1-4503-2598-1/14/11.
http://dx.doi.org/10.1145/2661829.2661835.

of a general `organization`. Using this category hierarchy, we can conduct analyses for entire groups of entities, for example, comparing the presence of `environmental organizations` and `power companies` in news of different parts of the world, deriving a picture of how their importance changes over time. The hierarchical organization of categories adds another dimension for aggregation to the usual temporal and spatial dimensions (where news can be aggregated by publishing times and originating regions).

2. RELATED WORK

There is ample work on text analytics for identifying (and visualizing) trends and patterns. Although this is of importance also for enterprise documents, most of the published research addresses social media (see, e.g., [8] and references there). The goals here are manifold: discovering events [3], connecting topics and users [1], and more.

Applications of such methods include culturomics [9] over the Google books corpus and computational journalism [5]. Although some of this work refers to "entities", the granularity of analyses really is noun phrases, disregarding the ambiguity of entity names. Also, there is no awareness of background knowledge bases. Two notable exceptions are the recent works by [10] and [7]. The former is a proof-of-concept project for annotating Web archive contents with entities. The latter applies shallow methods for entity markup to the French newspaper Le Monde, to support entity-aware culturomics. Our system uses much deeper methods for entity disambiguation and supports much richer analytics over both entities and categories. Other related work addresses the retrieval of entities in web or enterprise search (e.g., [2, 4]), as opposed to retrieving documents.

3. NEWS ANALYTICS ARCHITECTURE

The AESTHETICS engine for news analytics allows users to quickly spot trends of entities or groups of entities specified by a semantic category. The trend visualization is computed on daily occurrences of entities in articles of nearly 300 news sources. Continuously gathering news since June 2013, the AESTHETICS corpus now comprises more than 1.1 million articles. In each article, all entities are discovered and disambiguated using AIDA [6], which links all mentions in the article to entities in YAGO. YAGO contains ca. 500,000 semantic categories, comprising 4 million individual entities for which users can query. The entity index of AESTHETICS contains ca. 300,000 distinct YAGO entities which have been spotted in the news, and ca. 22 million entity occurrences.

One challenge is to support the user when specifying the entities and categories of interest for the actual analysis. The scale of YAGO renders the naive approach of letting a user choose from a list of everything infeasible. We solve this problem by providing automatic suggestions for entities and categories while the user is typing in the search field. When selected, a suggestion becomes a tag in the query field. The suggestion for entities and categories use a similar approach, first finding potential matching candidates based on a prefix matching, then ranking the candidates appropriately:

Entity Suggestions. All canonical entity names are stored in a trie for fast prefix lookup. Additionally, all token-wise suffixes are stored. For this, the name is split at whitespaces into tokens, and the complete string starting at each token is added as an additional label for the entity in the trie. This allows users to start typing a name in the middle; for example when typing "Obama", AESTHETICS still returns `Barack Obama` as a candidate. All matching candidates are then ranked based on the overall popularity in Wikipedia, which is estimated by the number of incoming links. Thus, typing "Obama" results in the suggestion `Barack Obama`, and `Michelle Obama`, in that order.

Category Suggestions. As with entities, all category names are stored in a trie. In contrast to entity names, category names can have an arbitrary order of words. For example, both `United States presidents` and `Presidents of the United States` are perfectly valid categories, and it is up to the Wikipedia edtiors' judgment which one to use. To cope with this diversity issue, we add all 3-token permutations as additional lookup keys, making sure that the real category shows for any permutation of {"united", "states", "president"}. The ranking is done based on the popularity of the contained entities, preferring categories with high average popularity. However, when typing "compan", the expected result should not be `American IT companies`, even though their average popularity might be very high; instead our method prioritizes more generic `IT companies` or just `companies`. Thus, the popularity-based ranking is balanced with a preference for categories that are closer to the root of the class hierarchy the categories are organized in.

A screenshot of the AESTHETICS Web interface is shown in Figure 2. In the search box at the top, entities, categories, and strings can be specified using the auto-completion method described above. On execution of the query, our system displays a chart showing the trend lines based on the daily occurrences of each item in the central area.

4. DEMO SCENARIOS

Our AESTHETICS system provides rich functionality for ad-hoc search and analytics over more than 1.1 million news articles. Conference participants will be able to flexibly explore the system's capabilities. The following are some of the conceivable use-cases.

Accurate entity counts. Entity-based analysis improves the accuracy over string-based analysis. The precision is improved as ambiguous names are resolved to the correct entity, removing unwanted occurrences. Better recall is achieved by also finding occurrences of the same entity under different names. Figure 1 shows a graph contrasting the various ways of specifying the target of the frequency analysis, using the example of the Maidan square in Kiev. Searching merely for "Maidan" overestimates the actual numbers by a factor of up to 2 (in February), while sometimes underestimating the true count (in late January). Searching for its English name "Independence Square" is not useful, either, as there are squares with this name all around the world.

Entity co-occurrence analysis. Comparing individual entity occurrences is already a powerful way of analyzing news and spotting trends or interesting topic shifts. A common scenario is that there is one main entity of interest, e. g. the `Ukraine`, and an analyst is interested in viewing and interpreting news in terms of other entities co-occurring with the main entity. Consider again the events around the Ukrainian crisis, analyzing the involvement of `Barack Obama` and `Vladimir Putin` over time, as depicted in Figure 3. AESTHETICS provides the advanced functionality of viewing occurrences for the three entities `Crimea, Obama`, and `Putin` given that `Ukraine` has to be mentioned in the

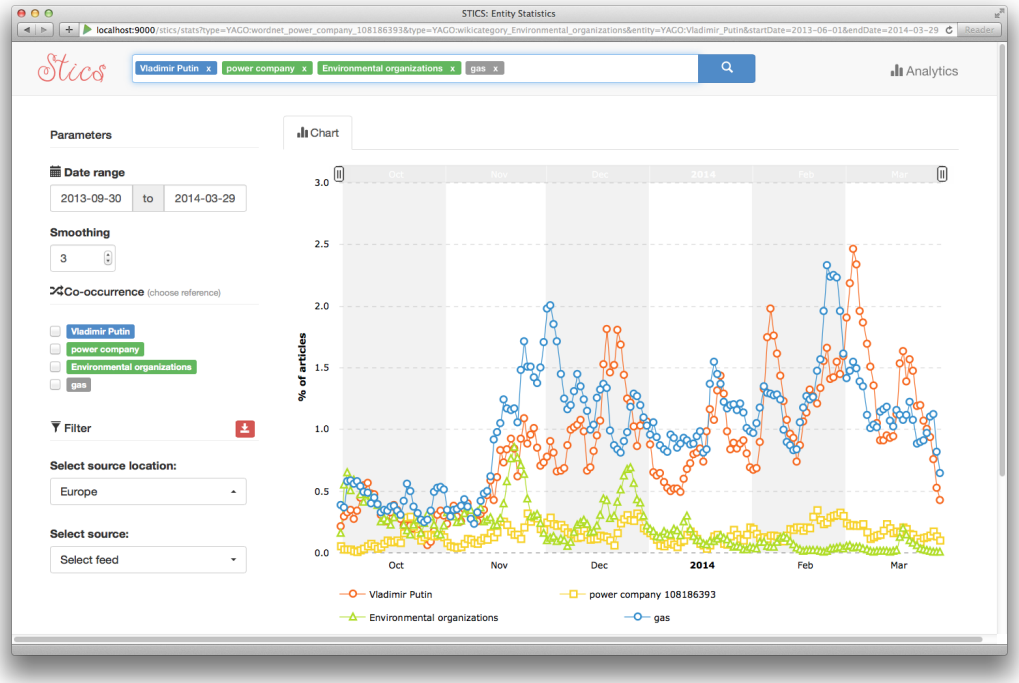

Figure 2: Analyzing the connection between `Vladimir Putin` and "gas" in the Aesthetics interface.

text. Without this constraint, analyzing the impact of the event on mentions of `Obama` is hard, as he is mentioned in a large number of totally unrelated contexts. Using the co-occurrence restriction, it becomes easy to spot that he became only involved once the `Crimea` situation escalated, and not in the initial protests in `Ukraine`.

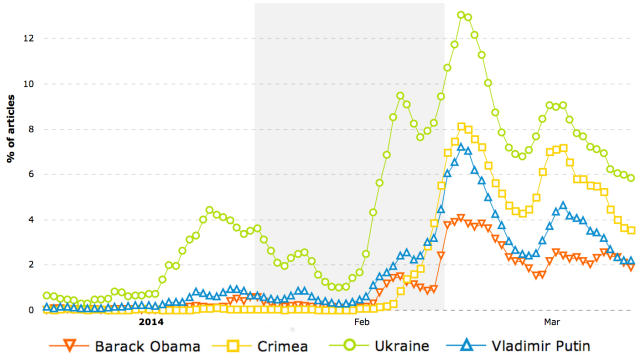

Figure 3: Entities in the context of the `Ukraine` crisis

Analyzing semantic groups of entities. Analyzing occurrences of all entities in a semantic group (e.g., Ukrainian politicians, female political activists, etc.) is impossible with just strings. With unstructured text linked to entities in a category hierarchy, it becomes a natural thing to do. The screenshot in in Figure 2 shows an example of an insightful analysis of this kind. The figure shows how often `power companies` or `environmental organizations` have been mentioned in news. This analysis reveals that the rise of the Crimea tensions in February had a remarkable effect on the frequency of mentions of power companies – driven by Europe's dependence on Russian gas, which is often transported through Ukraine. On the other hand, mentions of environmental organizations, opposing environmentally hazardous ways of getting gas by fracking, have gone down. This anecdotic evidence is strengthened by the observation that the effect is especially pronounced in European news, and harder to spot when removing the region filter.

Our demo is available online at
`http://www.mpi-inf.mpg.de/yago-naga/stics/`

5. REFERENCES

[1] S. Amer-Yahia et al.: MAQSA: A System for Social Analytics on News. SIGMOD 2012

[2] K. Balog, M. Bron, M. de Rijke: Query modeling for entity search based on terms, categories, and examples. TOIS 29(4), 2011

[3] A. Das Sarma, A. Jain, C. Yu: Dynamic relationship and event discovery. WSDM 2011

[4] H. Bast, B. Buchhold: An Index for Efficient Semantic Full-Text Search. CIKM 2013.

[5] A. Y. Halevy, S. McGregor: Data Management for Journalism. IEEE Data Eng. Bull. 35(3), 2012

[6] J. Hoffart et al.: Robust Disambiguation of Named Entities in Text. EMNLP 2011

[7] T. Huet, J. Biega, F. M. Suchanek: Mining History with Le Monde. AKBC 2013

[8] Sharad Mehrotra (Editor): IEEE-CS Data Engineering Bulletin 36(3), Special Issue on Social Media and Data Analysis, 2013.

[9] J.B. Michel et al. 2011. Quantitative Analysis of Culture Using Millions of Digitized Books. Science 331(6014), 2011

[10] M. Spaniol, G. Weikum: Tracking entities in web archives: the LAWA project. WWW 2012