Knowledge Linking for Online Statistics

Marc Spaniol, Natalia Prytkova, and Gerhard Weikum
Max-Planck-Institut für Informatik, Saarbrücken, GERMANY
{mspaniol|natalia|weikum}@mpi-inf.mpg.de

Abstract

The LAWA project investigates large-scale Web (archive) data along the temporal dimension. As a use case, we are studying Knowledge Linking for Online Statistics.

Statistic portals such as eurostat's "Statistics Explained" (http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Main_Page) provide a wealth of articles constituting an encyclopedia of European statistics. Together with its statistical glossary, the huge amount of numerical data comes with a well-defined thesaurus. However, this data is not directly at hands, when browsing Web data covering the topic. For instance, when reading news articles about the debate on renewable energy across Europe after the earthquake in Japan and the Fukushima accident, one would ideally be able to understand these discussions based on statistical evidence.

We believe that Internet contents, captured in Web archives and reflected and aggregated in the Wikipedia history, can be better understood when linked with online statistics. To this end, we aim at semantically enriching and analyzing Web (archive) data to narrow and ultimately bridge the gap between numerical statistics and textual media like news or online forums. The missing link and key to this goal is the discovery and analysis of entities and events in Web (archive) contents. This way, we can enrich Web pages, e.g. by a browser plug-in, with links to relevant statistics (e.g. eurostat pages). Raising data analytics to the entity-level also enables understanding the impact of societal events and their perception in different cultures and economies.

Keywords: Entity Disambiguation, Knowledge Base Alignment, Knowledge Linking, Online Statistics

1. Introduction

The constantly evolving Web reflects the evolution of society in the cyberspace. Web-preservation organizations like the Internet Archive[1] not only capture the history of digitally born contents, but also reflect the zeitgeist of different time periods for more than a decade. Web-content repositories and digitization projects open up a new range of analytical opportunities and challenges along the temporal dimension [Alonso et al. (2007)]. Studies reveal “culturomics” phenomena [Michel et al. (2011)], track the trustworthyness of memes over time (truthy[2]) or even investigate the Web's predictive power (such as time explorer[3] or recorded future[4]). Obviously, this longitudinal data is a potential gold mine for researchers like sociologists, politologists, media and market analysts, or statisticians. For instance, a statistician might be interested in tracking and analyzing public statements made by

[1] http://www.archive.org

[2] http://truthy.indiana.edu/

[3] http://fbmya01.barcelonamedia.org:8080/future/

[4] https://www.recordedfuture.com/

representatives of financial institutions such as International Monetary Fund (IMF), World Bank or European Central Bank (ECB), characterizing the evolution of their attitude towards the European sovereign debt crisis.

In order to understand the mutual dependencies between Web (archive) contents and the evolving societal knowledge it represents, contextual information such as online statistics are desirable. Statistic portals such as eurostat's "Statistics Explained" provide a wealth of articles constituting an encyclopedia of European statistics. Together with its statistical glossary, the huge amount of numerical data offers an abundance of contextual information. However, identifying the most suitable statistical document given a specific Web (archive) content is a non-trivial task. This is due to the different nature of textual Web contents and statistical documents. While Web contents usually describe an event mentioning concrete entities, such as people, organizations or locations, statistical articles are fairly abstract by covering a certain topic, e.g. "Renewable Energy Statistics", instead. As a consequence, standard approaches that "simply" create contextual information by matching keywords or keyphrases onto a thesaurus are of limited advantage given this setting. In order to overcome this shortcoming, a semantic exploitation of the content is required, which allows a proper alignment of Web contents with online statistics. To this end, we explore a hybrid approach that combines textual similarity measures with semantics captured in knowledge bases.

In this paper we present knowledge linking for online statistics. Our system called LILIANA (LIve LInking for online statistic ANAlytics) allows live linking of Web (archive) contents with online statistics, thus, providing contextual information. In order to identify the relevant statistical article(s), we raise Web (archive) contents to the entity-level so that we can align them with statistical categories of eurostat's "Statistics Explained". In addition to textual similarity measures, the relevant statistical articles can be dynamically interlinked and, thus, provide valuable contextual information. As a result, we intend to narrow and ultimately bridge the gap between numerical statistics and textual Web contents.

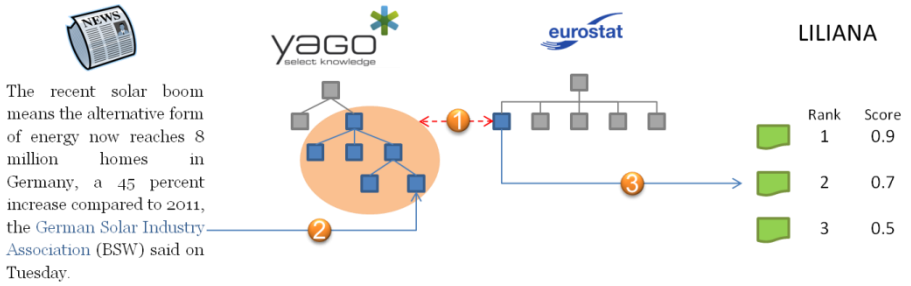

Figure 1: Pipeline of knowledge linking for online statistics

2. Conceptual Approach

In order to support knowledge linking for online statistics, LILIANA pursues a multi-stage procedure. This approach incorporates a semantic interpretation and linkage of Web contents. Since eurostat's "Statistics Explained" with its statistical glossary and its well-defined thesaurus represents a Wikipedia-like source, it can be interlinked with any other knowledge base such as Freebase [Bollacker et al. (2008)], Dbpedia [Auer et al. (2007)] or YAGO [Suchanek et al. (2007)]. As such, we first face an alignment problem on the ontological/entity-level. Then, we raise Web (archive) contents onto the entity-level. Finally, we perform live linking and ranking based on their textual and semantic similarity. By doing so, we provide contextual information from the relevant statistical article(s). Figure 1 depicts the knowledge linking pipeline by LILIANA, which we explain step-by-step in the following subsections.

Knowledge Base Alignment

We have chosen the YAGO knowledge base [Suchanek et al. (2007)] for semantic enrichment of textual Web (archive) contents. We obtained taxonomy, thesaurus and contents of "Statistics Explained" by crawling, thus, creating a replica for indexing and alignment. In particular, we have selected the "Statistical Themes" subsection of the hierarchy as it reflects a taxonomical structure used for classification.

Conceptually, both knowledge bases are organized in a Wikipedia-like structure. Therefore, we have undertaken in a first stage an alignment of YAGO and eurostat's "Statistics explained". However, they differ substantially in size and granularity. For instance, the "Statistical Themes" is fairly small and consists of only 40 categories in total used for classifying almost 2000 statistical articles (English contents only). On the contrary, YAGO contains almost 3 million entities classified by more than 350.000 types/categories derived from Wikipedia (cf. http://www.wikipedia.org) and WordNet [Fellbaum, C. (1998)]. Hence, there is no one-to-one correspondence between both knowledge bases.

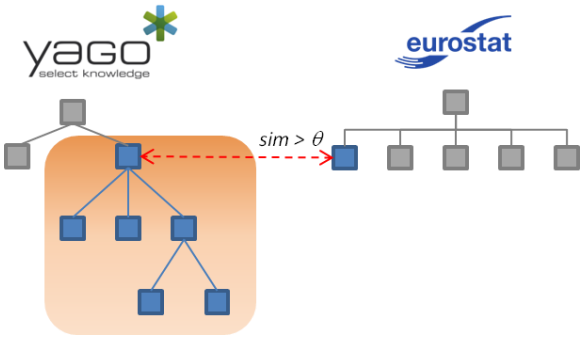

Figure 2: Knowledge base alignment

When aligning both knowledge bases, we started with those categories that can be directly mapped based on their textual similarity. This was do-able 12 of the 40 existing "Statistics Themes" categories. All other categories were then mapped onto their textual most similar counterpart, requiring that a user-definable similarity threshold θ is exceeded. In order to allow for the large discrepancy in size between the two knowledge bases, categories of "Statistical Themes" are mapped to a whole sub-hierarchy in YAGO (cf. Figure 2). Formally, the similarity between two categories c_1 and c_2 is defined as follows:

$$sim(c_1, c_2) = Jaccard(root(c_1), c_2) \cdot \frac{1}{distance(root(c_1), c_1)}$$

where $Jaccard(root(c_1), c_2)$ indicates the textual similarity between category c_2 and the root category of the subtree where c_1 is located. The $distance(root(c_1), c_1)$ shows how far c_1 is from its root in terms of edges. The mapping constructed in this way contains more than 1 million category pairs.

Entity Disambiguation for Semantic Enrichment

In order to semantically enrich textual Web contents, we lift the plain text to the entity-level by detecting named entities and resolving ambiguous names. To this end we employ the AIDA entity disambiguation system [Hoffart et al. (2011)] that maps mentions of entities onto canonical entities of the YAGO knowledge base [Suchanek et al. (2007)]. AIDA is built on top of the Stanford NER tagger [Finkel et al. (2005)] to identify mentions in the input document. As a result, we obtain a semantically enriched document including entities and their types. For instance, in a document that

contains the entity "European Central Bank"[5], we obtain 24 YAGO types it is associated with. Based on the previously described mapping we are able to interpret them indirectly as categories of "Statistical Themes" as well.

Live Linking and Ranking
We introduce in the following three models, which use semantic and textual features for linking and ranking.

TFIDF: Our baseline approach considers only the textual similarity between the query text and the documents in the corpus. Each document is represented as a *TFIDF* vector in the common feature space. The relevance of a document to the query is defined as a cosine similarity between the corresponding vectors:

$$score(a, q) = \cos(\overrightarrow{v_a}, \overrightarrow{v_q})$$

As outlined in the beginning, this method has several drawbacks. The constructing of a ranked list of recommended articles requires that the entire corpus has to be scanned and the cosine similarity to each document needs to be computed. Moreover, the result can be biased towards the most populated topics the collection.

Voting: This method works entirely on the entity-level**.** Here, the subset of the relevant articles is defined by means of the precomputed knowledge base alignment. Depending on types derived from the entities in the query text, the relevant documents in "Statistical Themes" are identified. The greater the overlap between entity and document types, the higher the document is scored. Let C be a set of entity classes mentioned in the query text q and C' their counterparts in "Statistical Themes". Then, we define the score of an article a as follows:

$$score(a, q) = |\{c' \in categories(a) | c' \in C'\}|$$

where $categories(a)$ return all categories of the article a. The voting model is fully semantic. However, if article classification is based on a coarse-grained type system (e.g. "Statistical Themes"), this commonly results in many equally ranked documents.

Voting + TFIDF: This approach is a combination of the previous methods in a two-stage computation. First, we confine the relevant statistical articles based on the semantic **voting** model. In a subsequent step, we then rank the documents based on the textual overlap of **TFIDF**. Thus, this model captures both, semantic and textual similarity. Formally, the score of an article a is defined as follows:

$$score(a, q) = |\{c' \in categories(a) | c' \in C'\}| \cdot \cos(\overrightarrow{v_a}, \overrightarrow{v_q})$$

3. Live Linking – Interconnecting Web (Archive) Contents and Online Statistics
Exploring and analyzing Web (archive) contents includes finding names of people, companies, products, songs, etc. Since names are often ambiguous, disambiguation of named entities in natural language text helps to map mentions onto canonical entities allowing a semantically exploitation Web data. However, raising mentions of people, places, or organizations onto the entity level is only the first step in making use of the raw and often noisy data. Even more, contextual information (e.g. online statistics) are required to understand the complex implications between Web (archive) contents and the underlying societal event being investigated.

Figure 3 shows the application of the LILIANA browser plug-in which is available for download at: https://addons.mozilla.org/de/firefox/addon/liliana-linking/). In the

[5] https://d5gate.ag5.mpi-sb.mpg.de/webyagospotlx/WebInterface?passedQuery=I%3A0%09S%3AEuropean\u005fCentral\u005fBank%09P%3Atype%09O%3A\u003fx%3B

example shown in the screenshot, the user has selected a text fragment of a news article dealing with "Rising Energy Prices – Germans Grow Wary of Switch to Renewables". Upon clicking on the right mouse button she is able to select the option "Link to Online Statistics". When doing so, the plug-in directs her to our disambiguation and link recommendation server (cf. left hand side of Figure 4) at: https://d5gate.ag5.mpi-sb.mpg.de/webliliana/. The user interface shown here, comprises the following four key components:

1) On top of the left panel, there are three buttons that can be selected. Depending on the user's choice, the underlying linking method is being selected. As a default setting, LILIANA employs the before mentioned hybrid approach that combines textual similarity measures with semantics captured in knowledge bases.
2) The text panel initially contains a copy of the text that has been selected when activating the LILIANA browser plug-in the previous step. However, the user may input any text, e.g. by copy-and-paste from arbitrary Web contents, or even HTML tables. By default, the AIDA entity disambiguation system identifies noun phrases that can be interpreted as entity mentions. As this is potentially error-prone, the user can alternatively flag mentions by putting them in double brackets, e.g.: "Harry is the opponent of [[you know who]]".
3) The output in the upper right pane shows for each mention (in blue), the entity that has been assigned to the mention, in the form of a clickable link. The links point to the corresponding Wikipedia articles. Alternatively, they could point to the YAGO knowledge base entries, or any comparable knowledge source in the Linked-Data world.
4) Finally, in the lower right pane links to the top ranked statistics articles are shown. In order to help the user in finding the most appropriate article, the title of the statistics article and the computed confidence score based on the selected linking method are shown.

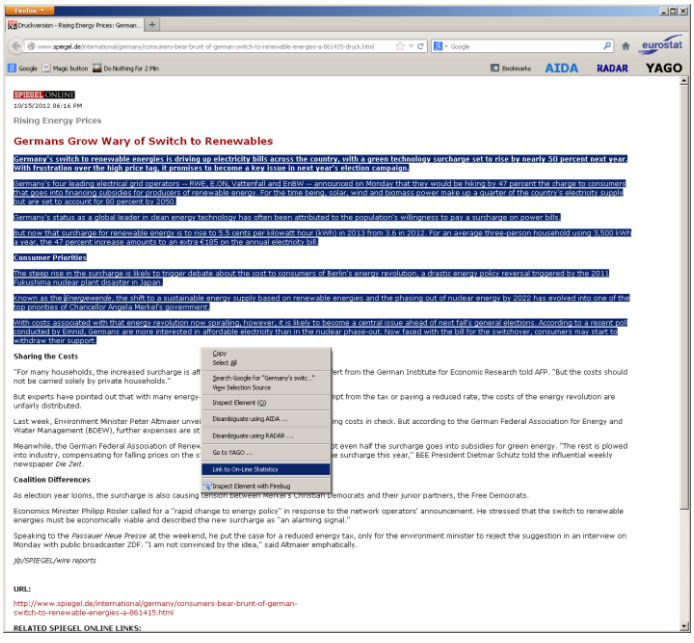

Figure 3: Live Linking by the LILIANA browser plug-in

The outcome of live linking for the highest ranked statistical article is shown on the right hand side of Figure 4. Based on entity and textual similarity, LILIANA points the user in this case to the article on "Renewable energy statistics" (cf. http://epp.eurostat.ec.europa.eu/statistics_explained/index.php/Renewable_energy_statistics).

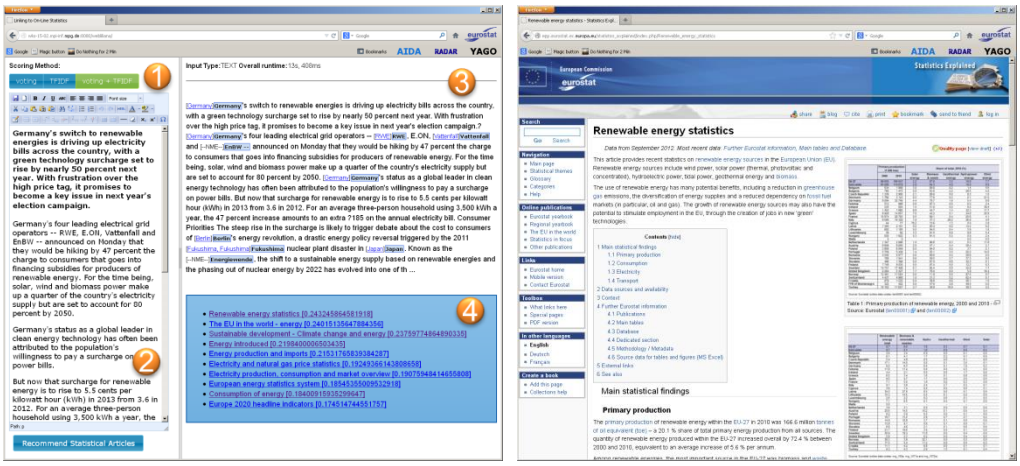

Figure 4: Entity-level analytics (left) and linked statistics article (right)

4. Conclusions

In this paper we have introduced the LILIANA system, which supports knowledge linking for online statistics. To the best of our knowledge, LILIANA is the first system that bridges the gap between numerical statistics and textual Web contents. As we believe that societal events and their perception in different cultures and economies captured in Web (archive) contents can be better understood based on contextual information, LILIANA provides links to the most related statistical article(s) by combining textual similarity measures with semantics captured in knowledge bases.

Acknowledgements

This work is supported by the 7th Framework IST programme of the European Union through the focused research project (STREP) on Longitudinal Analytics of Web Archive data (LAWA) – contract no. 258105.

References

Alonso O., Gertz M., and Baeza-Yates R. A. (2007) "On the value of temporal information in information retrieval". *SIGIR Forum*, 41(2):35–41.

Auer S., Bizer C., Kobilarov G., Lehmann J. Ives Z. (2007) "Dbpedia: A nucleus for a web of open data." *6th Intl. Semantic Web Conference*, *Springer*, pp. 11–15, Busan, Korea.

Bollacker K. D., Evans C., Paritosh P., Sturge T. and Taylor J. (2008) "Freebase: a collaboratively created graph database for structuring human knowledge". *SIGMOD*, pp. 1247–1250.

Fellbaum, C., editor (1998). "WordNet: An Electronic Lexical Database". *MIT Press*, Cambridge, MA.

Finkel J. R., Grenager T. and Manning C. (2005) "Incorporating Non-local Information into Information Extraction Systems by Gibbs Sampling", *ACL*, pp. 363–370.

Hoffart J., Yosef M. A., Bordino I., Fürstenau H., Pinkal M., Spaniol M., Taneva B., Thater S. and Weikum G. (2011) "Robust disambiguation of named entities in text." *Conference on Empirical Methods in Natural Language Processing (CoNLL)*, pp. 782–792, Edinburgh, Scotland, United Kingdom.

Michel J.-B., Shen Y. K., Aiden A. P., et al. (2011) "Quantitative analysis of culture using millions of digitized books". *Science*, New York, N.Y., 331(6014):176–182.

Suchanek F., Kasneci G. and Weikum G. (2007) "YAGO: A core of semantic knowledge – unifying WordNet and Wikipedia". *16th Intl. World Wide Web Conference (WWW 2007)*, *ACM*, pp. 697–706, Banff, Canada.