SESAME: European Statistics Explored via Semantic Alignment onto Wikipedia

Natalia Boldyrev[1] Marc Spaniol[2] Jannik Strötgen[1] Gerhard Weikum[1]

[1]Max Planck Institute for Informatics, Saarland Informatics Campus, Germany
[2]Université de Caen Normandie, Caen, France
{natalia,jannik.stroetgen,weikum}@mpi-inf.mpg.de
marc.spaniol@unicaen.fr

ABSTRACT

Authorities such as the European Commission have recognized the need to offer a unified access to the data gathered by a wide variety of providers, such as the European Statistical Organization (Eurostat) or the European Environment Agency. Its EU Open Data Portal[1] serves as a gateway to numerical data, statistical reports, and visualization tools. While making the data available to the users from all member states and concentrating efforts on bridging the language gap, the portal still focuses on a primarily statistical perspective. That is, numerical data are explained with general terms, only. However, the related events, people, or organizations "causing" or being "affected" by the statistical observation remain concealed to the user.

In order to make statistical data better understandable, we present the SESAME system (Statistics Explored via Semantic AlignMEnt). It relies on a novel method for identifying background information and relating it with event descriptions in Wikipedia. Using SESAME, users can jointly browse numerical statistics, their explanation in general terms and - now - also directly relate it to associated Wikipedia articles.

Keywords

Semantic Alignment; Linking Numerical Observations

1. INTRODUCTION

Motivation. With the abundance of emerging and continuously growing knowledge sources (e.g., in the linked open data cloud), a wide range of factual knowledge becomes available. Even more, highly specialized statistics gathered and maintained by professionals in governmental organizations such as Eurostat[2] serve as numerical evidences of facts. As shown in Figure 1, the reported number of asylum applications significantly increased in Germany in the year 2015. However, end users are often left alone with this information as background information on key concepts (e.g., events, people, or organizations) is missing.

In contrast, crowd-curated Wikipedia covers a wide range of concepts by detailed textual descriptions. For instance, consider the excerpts from the following Wikipedia articles:

> **Timeline of the European migrant crisis**
> "5 September 2015: German Chancellor, Angela Merkel, announced that there are 'no limits on the number of asylum seekers' Germany will take in."
>
> **Horst Seehofer**
> "In late 2015, Seehofer and the CSU sharply criticized Chancellor Angela Merkel's refugee policy."
>
> **Syrians in Germany**
> "During the European migrant crisis of 2014-2015 hundreds of thousands of Syrian refugees of the Syrian Civil War entered Germany to seek refugee status."

Obviously, those text snippets associated with events and persons are highly relevant to the observed statistical incline in asylum seekers in Germany in the year 2015. In addition to plain numerical data, these excerpts may provide valuable insights to better understand the overall context. However, identifying the relevant excerpts from free text is a non-trivial task as there is a low similarity between general statistical terms and the text itself.

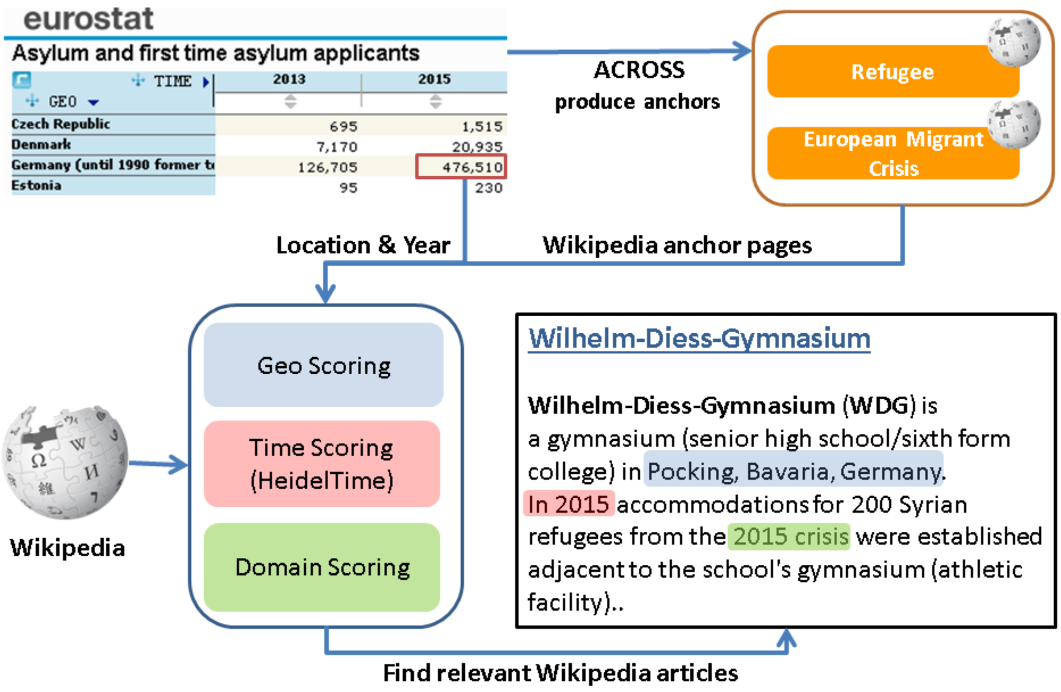
In order to unlock this hidden potential and successfully interlink numerical statistics with the relevant Wikipedia articles, we need to contextualize further dimensions, such as **time**, **location**, and **domain**. To this end, we introduce SESAME (Statistics Explored via Semantic AlignMEnt). SESAME bridges the gap between numerical statistics and semantically related Wikipedia articles by incorporating Wikipedia's link graph, temporal expressions within Wikipedia pages as well as page edit activities and views.

This presentation paper makes the following contributions:

- Semantic interlinking between numerical statistics and Wikipedia articles.
- Ranking of related Wikipedia articles based on their proximity to the time, location and domain of the desired statistical observation.
- A graphical user interface for jointly exploring numerical statistics and associated Wikipedia articles.

[1] http://data.europa.eu/euodp/en/data/
[2] http://ec.europa.eu/eurostat/web/regions/data/database

WWW'17 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3054732
.



GEO \ TIME	2013	2015
Czech Republic	695	1,515
Denmark	7,170	20,935
Germany (until 1990 former t	126,705	476,510
Estonia	95	230

Wilhelm-Diess-Gymnasium

Wilhelm-Diess-Gymnasium (**WDG**) is a gymnasium (senior high school/sixth form college) in Pocking, Bavaria, Germany. In 2015 accommodations for 200 Syrian refugees from the 2015 crisis were established adjacent to the school's gymnasium (athletic facility)..

Figure 1: Conceptual approach of SESAME.

2. OVERVIEW OF SESAME

SESAME consists of two conceptual building blocks: a mapping of statistical tables onto Wikipedia anchors and a contextualization module incorporating time, location, and domain information.

The captions of statistical tables are limited in the vocabulary and use very specific terms, for example, `First time asylum applicants` or `Primary production of renewable energy`. In the first step, SESAME associates each Eurostat table with a small number of Wikipedia pages, the so-called "anchors". The anchor pages define the scope of the table and are used for the subsequent contextualization step. For instance, the table `First time asylum applicants` is mapped to Wikipedia anchors `Dublin regulations` and `Right of Asylum`.

In the second step, SESAME now retrieves contextualized Wikipedia articles for the numeric statistics. To this end, we exploit the dimensions time, location and domain of the Wikipedia articles for scoring and ranking.

2.1 Mapping Tables to Wikipedia Anchors

Eurostat tables are referred to in statistical reports and used in them as evidences when describing changes and trends. This is illustrated in the paragraph below:

> **Asylum quarterly report**[3]
> "The number of first time asylum applicants in the EU-28 decreased by -15% in the third quarter of 2016 ...(Table 1)."

Eurostat organizes reports in a Wikipedia-like manner. Related reports are interlinked and appropriately categorized. Apart from outgoing links to other sources, a report points to the pages of the dedicated glossary section (e.g., `first time asylum applicants` in the example). This allows us to gather for each table a set of glossary entries.

Using ACROSS [2], glossary entries are converted to corresponding Wikipedia articles (anchors). Each table thus is interlinked with a set of anchors. In order to focus onto semantically coherent table-anchor assignments, we run ACROSS by applying the following constraint-aware reasoning:

1. a table is associated with a pair of non-correlating Wikipedia articles. A mapping of a table to both anchors - `Renewable Energy` and `Population` - is penalized. Two Wikipedia articles are said to be non-correlating, if the Jaccard similarity over their outgoing links sets is below a predefined threshold τ.

2. a Wikipedia anchor article annotates a pair of non-correlating tables. Two tables correlate positively, if they are mentioned in the reports of the same category. Both tables, `Asylum applicants` and `First residents permits`, belong to the category `Asylum and migration` and, thus, correlate.

2.2 Contextualization of Wikipedia Articles for Statistical Observations

Eurostat tables are two-dimensional matrices with rows being indexed by countries and columns by years. A statistical **observation** is a cell in this matrix. Together with the table annotations, an observation forms a query $\langle time,\ location,\ \{anchors\}\rangle$, where *time* and *location* are coordinates from the statistical table and *anchors* are the associated Wikipedia pages found in the previous step. In order to make observations in statistical data better understandable, our main goal is to provide meaningful background information to the user. To this end, we contextualize

[3] `http://ec.europa.eu/eurostat/statistics-explained/index.php/Asylum_quarterly_report`

each Wikipedia article w along the following three dimensions/features:

1) Temporal mentions. Using a temporal tagger, we determine all temporal expressions mentioned in w. Due to the largely narrative structure of Wikipedia, we use the domain-sensitive temporal tagger HeidelTime [9] with its narrative normalization strategy to correctly normalize not only explicit dates (e.g., `April 2002`), but also relative and underspecified expressions (e.g., `one month later` and `April`, respectively). As the temporal tag of an observation is always of year granularity, all extracted date expressions of finer granularities (e.g., `April 2002`) are mapped to the respective year (e.g., `2002`) and coarser expressions are ignored (e.g., `20th century`). Thus, each article a is associated with a multiset of year references.

2) Location mentions. In order to derive the set of location mentions of w, we consider all the outgoing links and treat them as entities. Using YAGO knowledge base [8], each entity is resolved to a semantic type and only those mapped to `yagoGeoEntity` are selected. Since the geo tag of the table observation is coarse-grained and is always a country, all location mentions in w are mapped to countries via the `locatedIn` relation. Thus, both locations `Berlin` and `Black Forest` are converted to `Germany`.

3) Outgoing links. The scope of w is determined by the set of its outgoing links.

Scoring. The introduced features are used to compute three proximity measures of a Wikipedia article w with respect to an observation $O = \langle time, location, \{anchors\}\rangle$. w is said to be relevant for O if:

1. w is semantically related to the *anchors*. Semantic similarity between page w and an anchor page a is defined as Jaccard similarity over their sets of outgoing links. The **domain relevance** is:

$$d(w) = \max_{a \in \{anchors\}} Jaccard(w, a)$$

2. w is semantically related to the *location*. Let G be the set of links pointing from w to any geo entity, and G' be the set of outgoing links to entities associated with *location*. The **location relevance** of page w is

$$g(w) = \frac{|G'|}{|G|}$$

3. w is relevant to the *time*. Let Y' be the number of year mentions related to the year of the observation and Y be the total number of temporal expressions. The **time relevance** for page w is

$$t(w) = \frac{Y'}{Y}$$

The final relevance score of w considers the three previously introduced proximity measures by a linear combination of their weights as follows:

$$rel(w, O) = \alpha \cdot d(w) + \beta \cdot g(w) + \gamma \cdot t(w) \qquad (1)$$

The time relevance can be further adjusted by considering the creation date of w. When looking for the recently emerged entities and events, the set of relevant articles can be further focused onto those, which were created in the year of the observation. However, since the Wikipedia history begins in January 2001 further constraining to earlier years is not possible for the given data set. Thus, using the creation time stamp of w is left to the user as an option, rather than a part of the scoring scheme.

3. DEMONSTRATION

3.1 Data

We have crawled the statistical reports from the Statistics Explained portal of Eurostat[4] in March 2016. The data contains 2,472 reports, 1,990 glossary terms and 557 categories. The tables with numerical statistics are taken from the Eurostat's Database[5]. In total, there are 2,398 tables which are mentioned in statistical reports. A minor fraction of the tables have time series of monthly or quarterly granularity. For the demonstration purpose and the sake of comparability, only tables with yearly statistics are exhibited to the user.

To compute the domain similarity of the pages, we use the static link graph derived from the English Wikipedia dump as of June 1, 2016. The revision history is parsed from the meta-history dumps of the same date and captures user activities starting from January 16, 2001. The page view data is retrieved from `http://stats.grok.se`.

3.2 Implementation

The interface is built using the Ace admin template[6], which is based on the Bootstrap framework and JQuery. On the server side, SESAME is written in Java and runs on a Tomcat server.

SESAME precomputes and stores the following data in a PostgreSQL database: (a) Wikipedia anchors for statistical tables, (b) the Wikipedia link graph, (c) annotation of Wikipedia articles with YAGO geo entities, and (d) temporal mentions found in Wikipedia articles by HeidelTime.

Through the Web interface, a user submits a query (table, location, year, weights for the three similarity measures) to the back-end engine. Each of the three scorers (see Fig. 1) retrieves relevant documents. To this end, the scores are aggregated using according to the user preferences. The scorers also return the "provenance" - the links through which the articles were considered to be relevant. These include location mentions and links to Wikipedia anchor pages. The server returns the ranked list of Wikipedia articles and the visualization interface renders the search results with provenance highlighting.

The SESAME user interface is divided into three parts.

1) Query box. The form on the left side serves for issuing the query. The user specifies the scope of numerical statistics to be contextualized by selecting a table, location and the year. Three types of weights - location, time and domain - may be adjusted by the user and are used for producing the aggregated score for relevant Wikipedia articles.

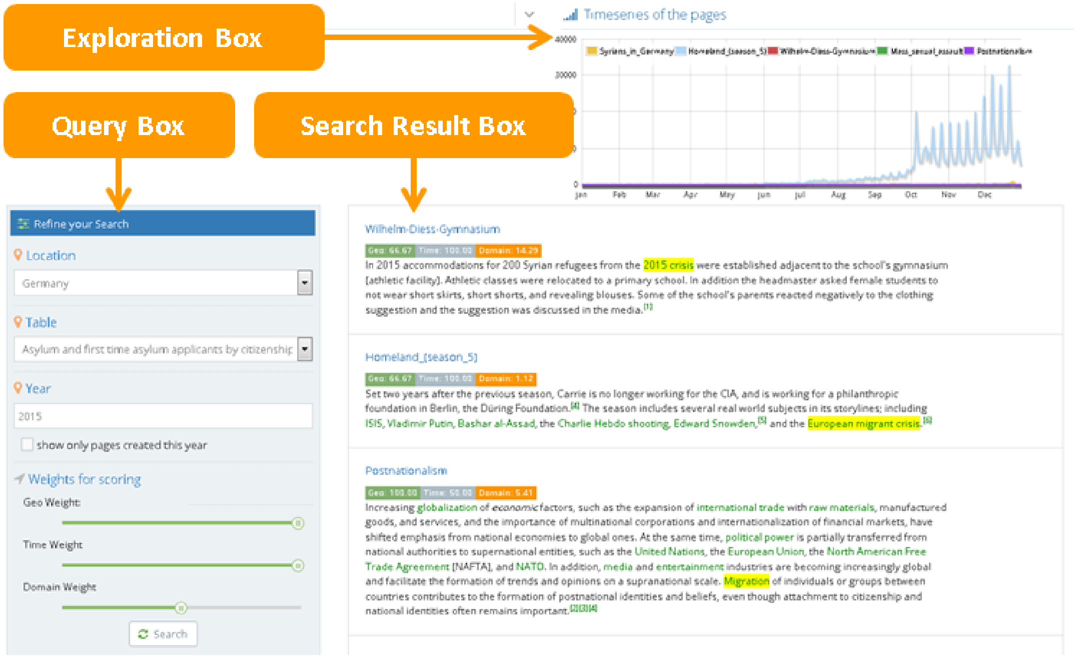
2) Search result box. After submitting the query, ranked search results are displayed on the right panel. Each article is represented in an expandable box. The box contains text snippets, where the anchors and relevant location mentions are highlighted.

3) Exploration box. The top panel provides two containers for further exploration. The left top box displays the

[4] `http://ec.europa.eu/eurostat/statistics-explained/index.php/Main_Page`

[5] `http://ec.europa.eu/eurostat/data/database`

[6] `http://ace.jeka.by/`



Location: Germany

Table: Asylum and first time asylum applicants by citizenship

Year: 2015

☐ show only pages created this year

Weights for scoring

Geo Weight

Time Weight

Domain Weight

Wilhelm-Diess-Gymnasium

Geo: 66.67 Time: 100.00 Domain: 14.29

In 2015 accommodations for 200 Syrian refugees from the 2015 crisis were established adjacent to the school's gymnasium (athletic facility). Athletic classes were relocated to a primary school. In addition the headmaster asked female students to not wear short skirts, short shorts, and revealing blouses. Some of the school's parents reacted negatively to the clothing suggestion and the suggestion was discussed in the media.[1]

Homeland_(season_5)

Geo: 66.67 Time: 100.00 Domain: 1.12

Set two years after the previous season, Carrie is no longer working for the CIA, and is working for a philanthropic foundation in Berlin, the Düring Foundation.[4] The season includes several real world subjects in its storylines; including ISIS, Vladimir Putin, Bashar al-Assad, the Charlie Hebdo shooting, Edward Snowden,[5] and the European migrant crisis.[6]

Postnationalism

Geo: 100.00 Time: 50.00 Domain: 5.41

Increasing globalization of *economic* factors, such as the expansion of international trade with raw materials, manufactured goods, and services, and the importance of multinational corporations and internationalization of financial markets, have shifted emphasis from national economies to global ones. At the same time, political power is partially transferred from national authorities to supernational entities, such as the United Nations, the European Union, the North American Free Trade Agreement (NAFTA), and NATO. In addition, media and entertainment industries are becoming increasingly global and facilitate the formation of trends and opinions on a supranational scale. Migration of individuals or groups between countries contributes to the formation of postnational identities and beliefs, even though attachment to citizenship and national identities often remains important.[2][3][4]

Figure 2: Overview of the SESAME exploratory interface.

selected Eurostat table and allows users to inspect the numerical statistics itself. The right box serves for highlighting the dynamics of the corresponding Wikipedia pages. To this end, it allows three types of user activities to be visualized: number of page views, number of editors, and number of page revisions aggregated per day for the query year.

3.3 Demonstration Scenario

Our demo covers an entire SESAME walk-through. Suppose, a user is interested in finding background information for the table `Asylum and first time asylum applicants`, for Germany in the year 2015. To build the query, the user navigates to the query box. The table and country are selected from the drop-down list and the year is typed into the corresponding field. Further, the feature weights can be adjusted with sliders according to the user's preferences.

Viewing the table data in the right exploration box reveals, that the number of asylum applications has risen from 202,645 in 2014 to 476,510 in the year 2015.

After submitting the query, the ranked Wikipedia articles are displayed in the search result box. Articles `Wilhelm-Diess-Gymnasium` and `Homeland (season 5)` are the top-ranked related pages, if location and time weight are set to 1.0 and the domain weight is 0.05.

> **Wilhelm-Diess-Gymnasium**
> "In 2015 accommodations for 200 Syrian refugees from the 2015 [migrant] crisis were established adjacent to the school's gymnasium (athletic facility)."
>
> **Homeland (season 5)**
> "The season includes several real world subjects in its storylines; including ISIS, Vladimir Putin, ... and the European migrant crisis."

The top right plot visualizes the page views and edit dynamics of the detected relevant articles.

Prominent articles such as `List of migrant vessel incidents in the Mediterranean Sea` attract considerable amount of community attention. On September 5th, 2015 this page had 1708 views. However, the long-tail article about Wilhelm-Diess-Gymnasium has a small number of editors and is also under-explored by Wikipedia users. It had moderate number of page views in 2015, with its maximum of 31 on September 9th and with at most 14 views per day until the end of 2015. Sesame is able to detect related pages independent of their prominence and unlock the hidden information from both, popular and long-tail articles.

The screencast of SESAME is available at:
`https://youtu.be/H2TiSTwqUhU`
The live demo of SESAME is available at:
`https://gate.d5.mpi-inf.mpg.de/sesame/`

4. EXPERIMENTAL EVALUATION

To evaluate the quality of SESAME alignments, we randomly selected 20 observations. We compare the following methods. SESAME with α, β, γ set to 0.5 (Formula 1). In these runs, the time relevance is controlled solely by parameter γ. SESAME + time additionally considers the creation date of Wikipedia articles (equivalent to checking `show only pages created this year` in the GUI). As baseline approaches, we choose two search engines – Bing and Google. To find relevant pages, we explicitly set the desired domain as `site:en.wikipedia.org` and formulate the queries as table name, location and year separated by a white space. A sample query might look as `Asylum and first time asylum applicants Germany 2015`. Further, we exploit the time settings utility provided by the search engines by additionally specifying the year of document creation (e.g., in Bing by setting `Date` → `Custom range` to 01.01.2015 to 31.12.2015 for the year 2015). These configurations are referred to as Google + time and Bing + time, respectively.

Method	MRR	MAP@10	Prec@10	Succ@1
SESAME	0.69	0.44	0.48	0.55
SESAME + time	0.54	0.17	0.27	0.41
Google	0.24	0.18	0.23	0.17
Google + time	0.20	0.03	0.13	0.0
Bing	0.41	0.11	0.23	0.22
Bing + time	0.14	0.02	0.05	0.0

Table 1: Experimental results.

Recommended Wikipedia pages are annotated by human judges either as *relevant* or *not relevant*. Our evaluation instructions stated that a page is considered to be relevant, if it is topically related to the table, as well as to the location and time.

The experimental results are given in Table 1. We report on four measures which are standard in information retrieval. Succ@1 refers to the portion of tables for which a relevant Wikipedia article was found at rank 1.

SESAME finds related Wikipedia pages with fairly high MRR and MAP@10 values. In contrast to SESAME, which treats locations as entities and also utilizes an underlying knowledge base to capture all locations belonging to a country by considering `locatedIn` relations, the search engines were not able to properly find pages related to the locations and time, treating these terms rather as keywords. This resulted in finding many general pages, which are location- and time-neutral (e.g., `Music education` or `Learning`).

Moreover, the table names are lengthy and are rather hard to deal with for a search engine. SESAME resolves this shortcoming by "reformulating" table names as a set of anchor pages. All the systems under consideration performed with lower MAP values when the creation date of Wikipedia pages was constrained (+ time option). This can be explained by the following observations: once an event has happened, there are many already existing Wikipedia pages which get updated (such as `Wilhelm-Diess-Gymnasium` in conjunction with the migrant crisis); pages for re-occurring events are created prior to their planned dates (already existing page `Olympics 2020` is a good illustration). By limiting the page creation date, all the systems loose a large portion of relevant results.

5. RELATED WORK

SESAME is related to several areas of previous research.

Semantic linking of tables [1], [3] aims at identifying entities mentioned in Web tables and the relations between them. The lack of a common scheme and high ambiguity of entity mentions are primarily addressed problems in this work. SESAME, in contrast, deals with tables having unified scheme. The focus of our work is to identify related documents from Wikipedia using time and location information as a part of ranking procedure.

Linking Wikipedia to external sources has been addressed in [4], [5], [6], [7]. The main focus has been done on extracting named entities, temporal expressions, and text excerpts from news articles, which make it a basis for interlinking with Wikipedia. SESAME, however, joins tables from Eurostat based on their semantic labels and performs a reasoning step to clean the alignment. Moreover, we consider the dynamics of Wikipedia page edits to rank these pages higher than background articles with low level of interest.

6. FUTURE WORK

SESAME has been successfully deployed with data from Eurostat. As a next step, we consider to incorporate additional statistics from the IMF, OECD, or UN.

7. CREATORS

Natalia Boldyrev is a PhD student at Max Planck Institue for Informatics. She obtained her MSc degree from Saarland University, Germany. Her area of research is alignment of heterogeneous knowledge repositories.

Marc Spaniol is a full professor at University of Caen Normandie, France. He is co-organizer of the Temporal Web Analytics Workshop (TempWeb) series. His research interests are in the area in the field of Web science, Web data quality, temporal Web analytics and knowledge evolution.

Jannik Strötgen is a postodctoral researcher at the Max Planck Institute for Informatics in Saarbrücken, Germany. He is the lead researcher of the multilingual, domain-sensitive tool HeidelTime and his research interests are in natural language processing and information retrieval.

Gerhard Weikum is leading the department on databases and information systems at the Max Planck Institute for Informatics in Saarbrücken, Germany. His research spans transactional and distributed systems, self-tuning database systems, data and text integration, and the automatic construction of knowledge bases.

8. REFERENCES

[1] C. S. Bhagavatula, T. Noraset and D. Downey. TabEL:Entity Linking in Web Tables. In *Proc. of ISWC*, 2015.

[2] N. Boldyrev, M. Spaniol and G. Weikum. ACROSS: A Framework for Multi-Cultural Interlinking of Web Taxonomies. In *Proc. of WebSci*, 2016.

[3] G. Limaye, S. Sarawagi and S. Chakrabarti. Annotatingand Searching Web Tables Using Entities, Types and Relationships. In *Proc. of VLDB*, 2010.

[4] A. Mishra, D. Milchevski and K. Berberich. Linking Wikipedia Events to Past News. In *Proc. of TAIA*, 2014.

[5] F. Nanni, S. P. Ponzetto and L. Dietz. Entity Relatedness for Retrospective Analyses of Global Events. In *Proc. of NLP+CSS*, 2016.

[6] M. Spaniol, N. Prytkova and G. Weikum. Knowledge Linking for Online Statistics. In *Proc. of WSC*, 2013.

[7] A. Spitz and M. Gertz. Terms over LOAD: Leveraging Named Entities for Cross-Document Extraction and Summarization of Events. In *Proc. of SIGIR*, 2016.

[8] F. M. Suchanek, G. Kasneci and G. Weikum. Yago - A Core of Semantic Knowledge. In *Proc. of WWW*, 2007.

[9] J. Strötgen and M. Gertz. Multilingual and Cross-domain Temporal Tagging. In *Language Resources and Evaluation*, 47(2), 2013.