Mining History with Le Monde

Thomas Huet
Max Planck Institute
for Informatics
Saarbrücken, Germany
thomas.h@ens.fr

Joanna Biega
Max Planck Institute
for Informatics
Saarbrücken, Germany
jbiega@mpi-inf.de

Fabian M. Suchanek
Max Planck Institute
for Informatics
Saarbrücken, Germany
suchanek@mpi-inf.de

ABSTRACT

The last decade has seen the rise of large knowledge bases, such as YAGO, DBpedia, Freebase, or NELL. In this paper, we show how this structured knowledge can help understand and mine trends in unstructured data. By combining YAGO with the archive of the French newspaper *Le Monde*, we can conduct analyses that would not be possible with word frequency statistics alone. We find indications about the increasing role that women play in politics, about the impact that the city of birth can have on a person's career, or about the average age of famous people in different professions.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

Keywords

YAGO, Le Monde, Culturomics, Knowledge Base, History, Newspaper, Mining

1. INTRODUCTION

Recent advances in information extraction have led to the creation of large knowledge bases (KBs). Projects such as YAGO [19], DBpedia [4], and Freebase[1], extract data from Wikipedia, and arrange it in a fact database. Other projects, such as NELL [7] or ReVerb [5], broadened this scope by extracting information also from unstructured sources, aiming at harvesting the entire Web. These KBs have accumulated billions of statements about millions of entities.

All of these projects strive to convert textual information into factual information. Their goal is to distill computer-understandable data from the vast amount of textual data that is out there on the Web. Once that information has been distilled, computers can use it for inference or query answering. Seen this way, human-produced natural language data helps computers structure and understand this world.

This endeavor has made huge progress in the last decade. Therefore, we believe that the time has come to turn that paradigm around: Computers have now accumulated so much structured data, that they should help *us* understand the unstructured information. We propose to use the factual data accumulated by machines to interpret the textual data produced by humans.

Some ground-breaking work in this direction has been made recently in the Culturomics project [15]. This project analyzed the Google Books corpus, a collection of n-grams extracted from 40 million books. By tracing the frequency of certain words over time in this collection, the project could show how language evolves, how certain words or concepts appear, how certain inventions become popular and then fade to unimportance, or how certain people rise to popularity. Thereby, the Culturomics project could identify trends in some of the vast amount of textual data that humanity has produced. This work, however, was based only on the frequency of occurrence of certain words. We believe that our understanding of massive textual data can be lifted to a new level if we tap into the large factual resources that machines have built up in the last decade. By using structured knowledge to interpret the unstructured data, we can gain new insights that go beyond the frequency of words.

To illustrate this point, we make use of the archive of *Le Monde*, the largest French newspaper. The archive reaches back to the year 1944, and covers the postwar period, the end of colonialism, and wars, politics, sports, culture, and economics until 1986. By linking the names in the articles to a KB, we can provide a deep analysis of trends during that period – deeper than would be possible by tracing word occurrences. The factual data of the KB allows us to trace developments, make statistical analyses, and gain deeper insights into the newspaper articles. By this analysis, we want to exemplify how structured data can help us understand unstructured data.

2. RELATED WORK

Our endeavor is related to a broad range of topics.

Entity Extraction. The recognition and classification of entity names in a text corpus has a long history. Statistical methods and rule-based methods are used to find entity names, and to classify them into people, locations, or organizations. We refer the reader to [18] for a recent survey of methods. Our proposed approach uses these techniques, but goes beyond them by illustrating the added value that the connection with a KB can bring to a text corpus.

[1] `http://freebase.com`

AKBC'13, October 27–28, 2013, San Francisco, CA, USA.

ACM 978-1-4503-2411-3/13/10
http://dx.doi.org/10.1145/2509558.2509567

Disambiguation. For our proposed approach, the entities extracted from the corpus need to be disambiguated. This task was addressed by many previous works. Some more recent ones are, e.g., [6], [8] and [12].

Knowledge Base Construction. Automated knowledge base construction is a relatively recent field of work, which has led to KBs such as YAGO [19], DBpedia [4], Freebase, TextRunner [5], or NELL [7]. Our proposal uses these KBs as an input. By analyzing the newspaper corpus, our approach provides value that the KBs by themselves cannot yield.

Ontology-based solutions. Knowledge bases have already been used to solve a number of different tasks. Those include e.g. information extraction [1], information structuring [9] and document clustering [13]. Moreover, some solutions have been proposed to mine patterns using a taxonomy of concepts extracted from text resources [14], [10]. These approaches, however, precede the arrival of large KBs and could thus not make use of them.

Event Prediction. Most of the closely related work focuses on the event prediction based on news corpora. Similarly to our approach, some methods use ontologies to assist the process of knowledge discovery. In [17], the authors use Linked Data to build entity features used by a probabilistic event predictor, as well as to generalize the entities. The use of an ontology in our work is similar. However, we propose to use the factual knowledge in order to compute aggregated statistics and, as a result, mine trends and patterns.

Culturomics and linguistic analyses. Based on Google's effort to digitize books, the Culturomics project [15] has mined cultural, social and linguistic trends in textual data over time. [3] introduces a notion of computational history – a discipline that might assist historians in analyzing the massive amount of information about the past. In the paper the authors focus specifically on how certain topics are remembered, examining different articles from the Google News Archive. In [16] the authors mine news corpora to discover semantic relatedness of words. [11] provides tools for creating heat maps of concepts as a response to user queries, visualizing concept relatedness.

Although related, the aforementioned projects did not make use of large KBs. With our work, we aim to show that these KBs can greatly assist the trend discovery process, and lead to insights that cannot be achieved by statistical analysis of words alone.

3. STRATEGY

Le Monde. The Google n-grams corpus provides a wealth of textual data. However, for our purpose, we are less interested in general trends in language usage, but rather in textual coverage of historical events. Therefore, we resorted to the archive of *Le Monde*, the largest French daily newspaper. This archive comprises all articles ever published in the newspaper, covering the years 1944-2013. However, due to the ongoing data curation at *Le Monde*, we had access only to the data for the years 1944-1986. This limits the conclusions that we can draw form the analyses, but does not stand in the way of the general point that we are making. The articles cover events in politics, culture, economics, and sports. Each article comes with a title, a publication date, and the full article text. In total, our corpus contains 502,781 articles, with a total size of 3 GB.

YAGO. Our goal is to illustrate the added value that a KB can bring to our understanding of textual data. There are many KBs that could be used for this purpose. For our work, we used YAGO [19], because it provides a good coverage of commonly known entities at a high accuracy. YAGO contains data from Wikipedia, WordNet, Geonames, and other sources. In total, the KB knows 10m entities and 100m facts about them.

Entity Recognition. For our analysis, we need a record of all entities that appear in the articles. There is an ample body of research about how to detect proper names and how to disambiguate them to entities in a KB (see Section 2), but most of these methods are expensive. For our work, we used a particularly simple and inexpensive approach, as it proved to work sufficiently well for our purposes. We first collected statistics on how often a particular link anchor text in the French Wikipedia refers to a particular French Wikipedia article. We excluded very infrequent phrases. The French Wikipedia articles can be mapped to YAGO entities by the multilingual information in YAGO. We then scanned the whole corpus for phrases that we could map to YAGO entities, excluding stop-words and lower-case phrases. This approach suffers from the incompleteness of both YAGO and the French Wikipedia, but we assume that the most important entities (which are also the ones that are more commonly mentioned in the articles) appear in both. In total, we were able to extract 3,425,656 entity mentions. A manual evaluation on a sample showed that these mentions are mapped correctly to YAGO with a precision of 86.8%, and that we achieved a recall of 77.1%. This phase yields a table that contains, for every article, each YAGO entity that appears in it.

Location Detection. Each article in *Le Monde* usually describes an event that takes place in some country. [2] describes a technique to detect the geographical location where an event took place. We used a simplified version of this technique as we were only interested in granularity at the level of countries: The country chosen for an article is the one that is mentioned most often in this article, with a mention of a place inside the country counting as a mention of the country itself. This analysis allows us to construct a table that contains, for every article, the location of the event described in the article. This table, together with the table of entity mentions and the publication dates of the articles, is all we require for our task.

4. RESULTS

In order to show how different aspects of a KB can assist in the knowledge harvesting process, we have constructed and performed a number of analyses along several dimensions.

Temporal trend mining. The Culturomics project mined the changing gender balance in the Google Book corpus. However, since the project did not have a KB at its disposal, it could only mine for occurrences of the words "men" and "women". With a KB at our hands, we can make that analysis much more detailed. By the help of YAGO, we mined the occurences of all female politicians, male politicians, females, and males that YAGO knows of. For each year, we calculated the number of articles that mention an entity from a given group and then normalized the mention count by the number of articles published that year. The results are shown in Figure 1. Although one can see a slight growth in the newspaper presence of female politicians and

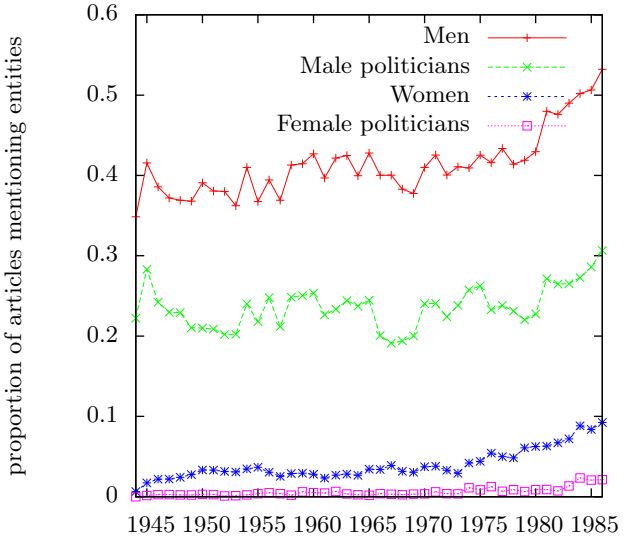

Figure 1: Mentions of men and women over time.

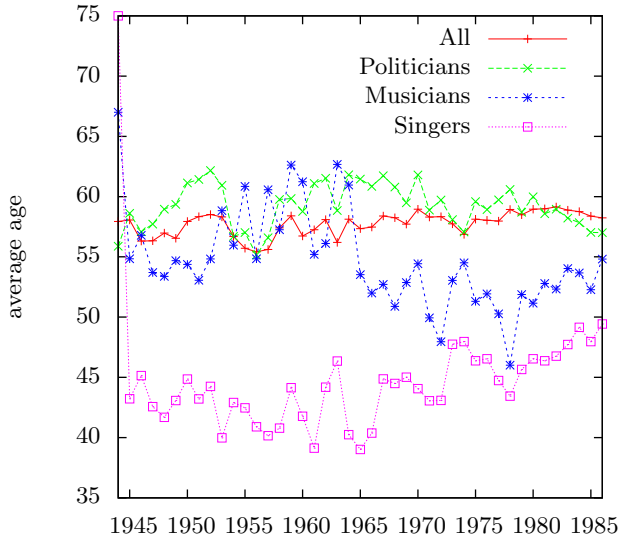

Figure 2: Average age of people mentioned.

females in general, there is still a big gap. This analysis is much more detailed than a search for the words "men" and "women", because it is based on the mentions of actual persons. It also allows slicing and dicing by professions or other criteria.

In a similar manner, we mined the occurences of diseases. Preliminary and yet-to-be-confirmed results show a steady increase in the mentions of diseases in France (despite an overall growing life expectancy).

In these examples, the KB provided us with the information about the types (politician, person, disease), the locations (France, elsewhere) and the gender of people (male, female). It was thus possible to mine trends concerning whole groups of entities. Simple keyword analysis does not reach this level of generalization.

Temporal trend mining with aggregations. We can employ numeric data to compute different statistics about the entities occuring in the corpus. As an example, we propose calculating the average age of the politicians, musicians, singers and of all people mentioned in our articles (Figure 2).

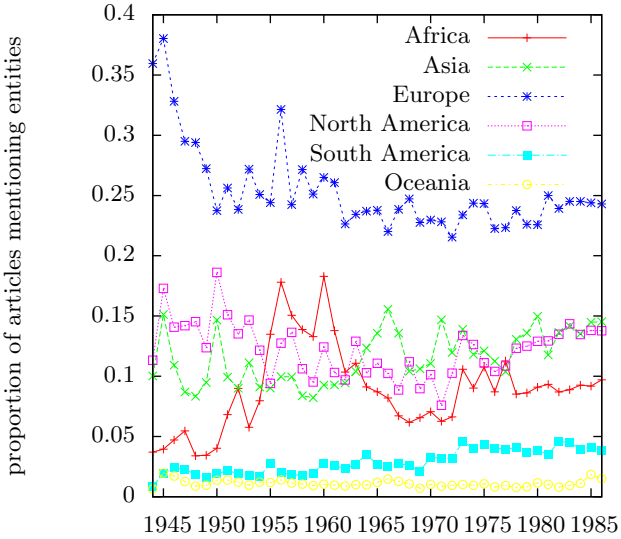

Figure 4: Mentions of countries per continent.

The results show that politicians mentioned tend to be older than the average while singers tend to be much younger. We restricted our analysis to contemporary people, i.e. aged between 20 and 100 at the time of the article publication. We calculated the average age over all the contemporary people mentions each year.

Here, we used not only the types and the locations of the entities, but also their birth dates. This enabled us to calculate the age of entities at the time of article publication – an analysis that goes beyond the frequency of names and shows the value of the KB for knowledge discovery.

Temporal anomaly mining. We believe that structured knowledge can also assist mining unusual patterns and, as a result, indicate important events. To visualize this, we mined the occurences of countries grouped by the continents (Figure 4). First of all, we can see a post-war decrease of mentions of the European countries. The most notable anomaly, however, stands out for the African countries around 1960, which is indeed the time when many of them gained independence.

The structured knowledge used here consists of the type information (country) and the location information (Europe, Africa, etc.).

Spatial statistics mining. To illustrate how the spatial knowledge can be used in mining textual corpora, we propose to mine the proportion of mentioned people who are born in the capital of their country. The results are shown in Figure 3, where red indicates a high proportion of people mentioned born in the capital and blue a low one (countries for which we have less than 42 mentions are shown in grey). This measure is high for Norway (98%) or France (87%) but low for China (8%) or the US (7%). This could provide an indication of how centralized the countries are, or whether the opportunities for people are distributed evenly.

Furthermore, using the geolocation of articles, we tested what percentage of foreign companies (i.e. not headquartered in a country) is mentioned in articles located in different countries. Results are shown in Figure 5, where red indicates a high proportion of foreign companies, and blue indicates a high proportion of domestic ones. This data has to be taken with a grain of salt, because *Le Monde* is focused

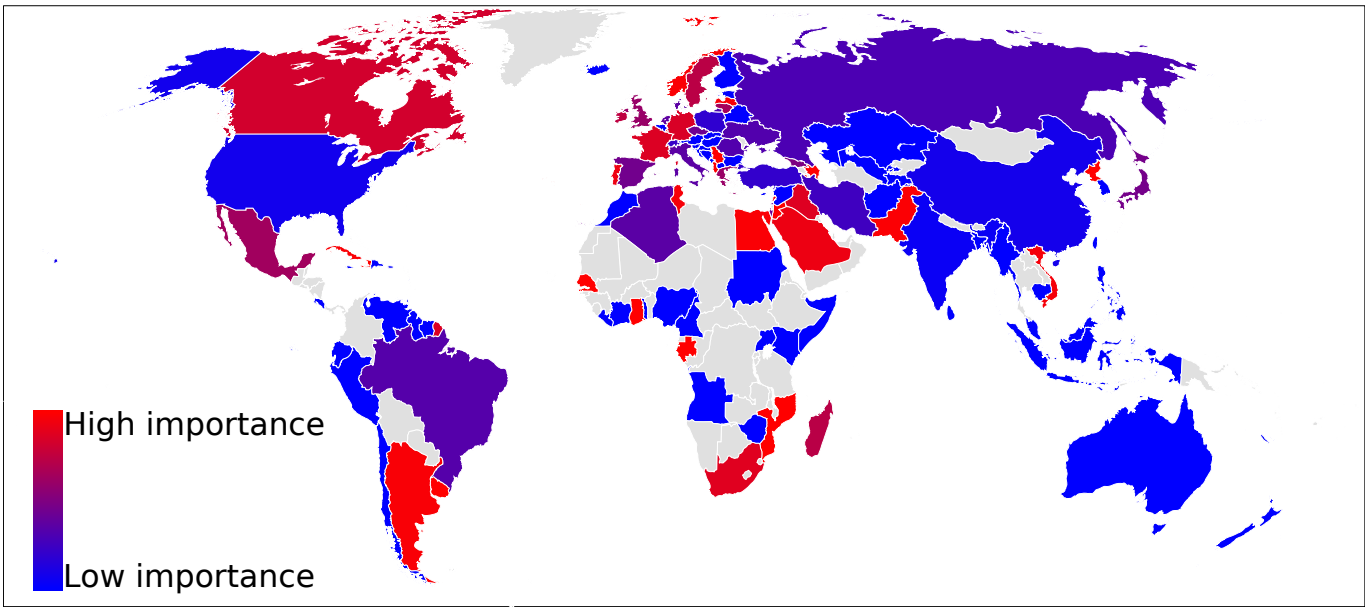

Figure 3: Importance of the capital per country

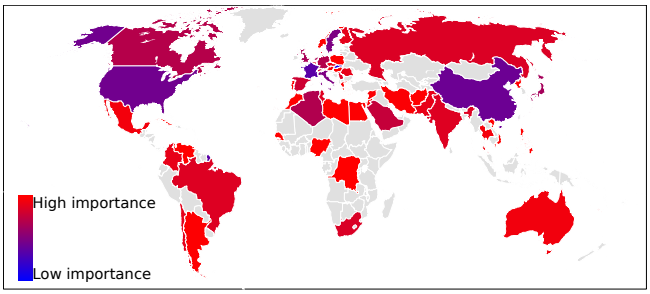

Figure 5: Importance of foreign companies

Headquarter country	Company mentions
France	18584
United States	10799
United Kingdom	2819
Germany	2033
Italy	1979

Figure 6: Company mentions

on French companies. As shown in Figure 6, French companies are mentioned in the corpus most often. Nevertheless, the analyses might indicate how big the impact of foreign subjects in certain countries is.

In these examples, the information drawn from YAGO included the types, locations, birth dates, birth places, and the capitals of countries. We also incorporated the geolocation information that we previously generated for each article.

Entity-oriented statistics mining. Additionally, ontological knowledge can be useful for creating entity rankings. For instance, we mined a list of globally present people by finding what are the entities that are mentioned in the largest number of locations. To rule out occasional occurences, we considered only the pairs of entities and locations that occur in the corpus at least 5 times. Figure 7 shows the top global persons of *Le Monde*. This idea, however, does not need to be restricted to people. In fact, one can create similar rankings for any subset of entities. Our experiments included, e.g., searching for globally present companies.

Person	Countries
François Mitterrand	94
Valéry Giscard d'Estaing	74
Ronald Reagan	71
Charles de Gaulle	68
Jesus	50
Pope John Paul II	47
Joseph Stalin	46
Nikita Khrushchev	45

Figure 7: Most globally mentioned people

In these examples, we made use of the entity type information and the article geolocation information.

5. CONCLUSION

In this paper, we have illustrated the added value that structured data can bring to the analysis of unstructured data. We wish to emphasize that our results describe the world as seen through *Le Monde* and YAGO, which may differ from the real world. Furthermore, both YAGO and our algorithms exhibit small imprecisions, so that we may not conclude without further investigation that our results would identify trends in *Le Monde*, let alone global trends. Rather, this paper made a proof of concept showing that the factual knowledge that machines have accumulated can now be used to interpret the textual resources that humans have produced. The structured data of a KB can enhance our understanding of corpora in a way that a keyword analysis alone cannot yield. We believe that, if this avenue of research is continued, it can lead to a better grasp of textual corpora – and possibly even to a better understanding of the events on this planet.

Acknowledgements. We would like to thank *Le Monde* for providing us generously with their archive.

6. REFERENCES

[1] H. Alani, S. Kim, D. Millard, M. Weal, W. Hall, P. Lewis, and N. Shadbolt. Automatic ontology-based knowledge extraction from web documents. *Intelligent Systems, IEEE*, 18(1), 2003.

[2] E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR*, 2004.

[3] C.-m. Au Yeung and A. Jatowt. Studying how the past is remembered: towards computational history through large scale text mining. In *CIKM*, 2011.

[4] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. G. Ives. DBpedia: A nucleus for a Web of open data. In *ISWC*, 2007.

[5] M. Banko, M. Cafarella, S. Soderland, M. Broadhead, and O. Etzioni. Open Information Extraction from the Web. In *IJCAI*, 2007.

[6] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proc. EACL*, 2006.

[7] A. Carlson, J. Betteridge, R. C. Wang, E. R. H. Jr., and T. M. Mitchell. Coupled semi-supervised learning for information extraction. In *WSDM*, 2010.

[8] S. Cucerzan. Large-scale named entity disambiguation based on wikipedia data. In *Proc. EMNLP*, 2007.

[9] D. W. Embley, D. M. Campbell, R. D. Smith, and S. W. Liddle. Ontology-based extraction and structuring of information from data-rich unstructured documents. In *Proceedings of the seventh international conference on Information and knowledge management*, CIKM '98, 1998.

[10] R. Feldman and H. Hirsh. Exploiting background information in knowledge discovery from text. *J. Intell. Inf. Syst.*, 9(1), July 1997.

[11] B. Hecht, S. H. Carton, M. Quaderi, J. Schöning, M. Raubal, D. Gergle, and D. Downey. Explanatory semantic relatedness and explicit spatialization for exploratory search. In *Proceedings of the 35th international ACM SIGIR conference on Research and development in information retrieval*, 2012.

[12] J. Hoffart, M. A. Yosef, I. Bordino, H. Fuerstenau, M. Pinkal, M. Spaniol, B. Taneva, S. Thater, and G. Weikum. Robust disambiguation of named entities in text. In *Proc. EMNLP*, 2011.

[13] A. Hotho, S. Staab, and G. Stumme. Ontologies improve text document clustering. In *ICDM*, 2003.

[14] S. Loh, L. K. Wives, and J. P. M. de Oliveira. Concept-based knowledge discovery in texts extracted from the web. *SIGKDD Explor. Newsl.*, 2(1), June 2000.

[15] J.-B. Michel, Y. K. Shen, A. P. Aiden, A. Veres, M. K. Gray, T. G. B. Team, J. P. Pickett, D. Hoiberg, D. Clancy, P. Norvig, J. Orwant, S. Pinker, M. A. Nowak, and E. L. Aiden. Quantitative analysis of culture using millions of digitized books. *Science*, 331(6014), 2011.

[16] K. Radinsky, E. Agichtein, E. Gabrilovich, and S. Markovitch. A word at a time: computing word relatedness using temporal semantic analysis. In *Proceedings of the 20th international conference on World wide web*, 2011.

[17] K. Radinsky and E. Horvitz. Mining the web to predict future events. In *WSDM*, 2013.

[18] S. Sarawagi. Information Extraction. *Foundations and Trends in Databases*, 2(1), 2008.

[19] F. M. Suchanek, G. Kasneci, and G. Weikum. YAGO: A core of semantic knowledge. Unifying WordNet and Wikipedia. In *WWW*, 2007.