How much is Wikipedia Lagging Behind News?

Besnik Fetahu, Abhijit Anand, Avishek Anand
L3S Research Center, Leibniz University of Hannover
Appelstr. 9a
30167 Hannover, Germany.
{fetahu,aanand,anand}@L3S.de

ABSTRACT

Wikipedia, rich in entities and events, is an invaluable resource for various knowledge harvesting, extraction and mining tasks. Numerous resources like DBpedia, YAGO and other knowledge bases are based on extracting entity and event based knowledge from it. Online news, on the other hand, is an authoritative and rich source for emerging entities, events and facts relating to existing entities. In this work, we study the creation of entities in Wikipedia with respect to news by studying how entity and event based information flows from news to Wikipedia.

We analyze the lag of Wikipedia (based on the revision history of the English Wikipedia) with 20 years of *The New York Times* dataset (NYT). We model and analyze the lag of entities and events, namely their first appearance in Wikipedia and in NYT, respectively. In our extensive experimental analysis, we find that almost 20% of the external references in entity pages are news articles encoding the importance of news to Wikipedia. Second, we observe that the entity-based lag follows a normal distribution with a high standard deviation, whereas the lag for news-based events is typically very low. Finally, we find that events are responsible for creation of emergent entities with as many as 12% of the entities mentioned in the event page are created after the creation of the event page.

Categories and Subject Descriptors

H1.1 [**Information Systems**]: Systems and Information Theory—*News and Wikipedia Dynamics*

Keywords

entity lag, event lag, news reference density, emergent entity density, wikipedia, news corpora

1. INTRODUCTION

Wikipedia is the largest source of open and collaboratively curated knowledge source in the world. Introduced in 2001, it has evolved to be a very useful repository for entities, events, concepts etc. Entities and event pages are often created and collaboratively edited creating a knowledge source which is both authentic and recent. As a result, this invaluable resource has found application in information extraction and knowledge base construction, e.g.YAGO [16] and DBpedia [4], text categorization [18], entity disambiguation [8], and entity ranking [10].

Owing to events being increasingly documented in online media, existing entities in Wikipedia continuously evolve and new entities are added. Moreover, online news has seen a lot of growth of late, and records important events reasonably quickly. Consequently, a high proportion of the entity pages in Wikipedia (pages devoted to named entities like persons, organizations and locations) have news articles as references, a factor which suggests that news is an authoritative source for important facts (we do a detailed analysis of the density in Section 4).

Automatic knowledge base construction tasks can rely on news as a source or an indicator to add or update entities. First, news could be a primary source for addition of emerging entities [6]. Secondly, knowledge bases which harvest this resource need to periodically refresh their contents. They constantly deal with the natural trade-off between the cost of maintenance of a fresh and consistent state with the loss of useful information. For newsworthy entities and events, understanding this delay in appearing in Wikipedia would suitably help knowledge bases to improve their maintenance or characterize the information loss.

In this work, we study how fast Wikipedia reacts to these real world events captured by news collections. We carry out our study on the Wikipedia revision history and the New York Times news corpus for the overlapping years between 2001 and 2007. We extract entities from the news articles and link them to the version of the entity page which is closest in time to the publication time of the article. In other words, we *align* the news collection to the Wikipedia versions using entities as proxies. Next, we define *lag* as the time difference between when an entity or event was reported in news and the first time it appeared in Wikipedia. This aligned resource allows us to carry out several studies which shed light on the evolution of entities and events and on how they are captured in Wikipedia.

Specifically, we try to answer the following questions:

- *What fraction of external references in entity pages are*

news articles ?

- *How much does Wikipedia lag behind news articles ? How has this lag evolved over time ?*

- *Which categories or classes of entities in news lead or lag Wikipedia ?*

- *How do events reported by news articles lag with the Wikipedia event pages ?*

We perform our alignment studies on the entire English Wikipedia revision history and the New York Times collection (as the news corpora). We also consider Wikipedia's current events portal as a repository for high quality manually created resource for events. Some of the highlights of our study are:

- Approximately 20% of all external references in entity pages are news articles.

- Entity lag follows a distinct normal distribution and show that Wikipedia has been catching up on news ever since it was introduced.

- Unlike entities, events are quickly reflected in Wikipedia as soon as they are reported in news.

- Events are responsible for creation of emergent entities, with 12% of the entities mentioned in event pages being created after the creation of the event page.

The rest of this paper unfolds as follows. In Section 3 we introduce the experimental setup and the generated data. Section 4 provides an overview of the news dynamics in Wikipedia and serves as a motivation to our study, while in Section 5 and 6 we provide thorough analysis and results on entity and event lag. In Section 2 we review related literature relevant to our work. Finally in Section 7 we conclude the findings in our analysis between Wikipedia and news corpora such as NYT.

2. RELATED WORK

The related work and state of the art with respect to our work can be classified into the following three parts :

Wikipedia Studies goes into similar directions with our analysis. Kittur et al. [13] analyses the collaborator structure of Wikipedia. They further classify the collaborators into five different classes based on the number of revisions. Furthermore, they measure the population growth of the collaborators falling into the five different classes. In their paper the authors conclude an interesting observation of the shift of how content is mostly provided by collaborators with lower number of edits, due to the increased fraction of such users in the Wikipedia community structure. This, however, does not correlate with any decline of the content provided by collaborators with high number of edits, hence, is accounted to the higher fraction of low edit users. In contrast to the work from Kittur et al., we have a different focus in our analysis, namely that of entity and event lag in Wikipedia, without any distinction of the Wikipedia community structure. In [17] the authors analyze several aspects of Wikipedia's editors. They conclude that the number of edits is decreasing. Another slightly related work [3] analyzes the number of research papers about Wikipedia, here too

they conclude that the number has been decreasing, however, papers that use Wikpedia's data has seen an increase.

Closely related work is done by Keegan et al. [11, 12, 9]. Their work, similarly to ours, focuses on the dynamics of Wikipedia's coverage of real world entities. In [11], the authors consider emerging events like the Tōhoku catastrophe[1]. In the case of such high dynamic events, it is found out that for localized Wikipedias (e.g. Japanese), the corresponding event appears only after six minutes after the event, whereas in the English Wikipedia, it appears in less than an hour. Furthermore, they analyze the co-authorship of such articles in Wikipedia. It is concluded that within Wikipedia there are sub-communities that edit articles of the same topic. As a continuation of their work, in [12] the social network structure of Wikipedia collaborators is analyzed. The analysis is based on four main hypotheses that are based on two main set of attributes, article and editor attributes, respectively. The first hypothesis validates the fact that for breaking news articles attract more editors. The second hypothesis validates the co-authorship of articles in Wikipedia from collaborators that are categorized into three main classes: *Experienced, Apprentice, Non-Expert.* Significant collaborations between the three classes of collaborators is found only on *contemporary* articles (articles are divided into *breaking, contemporary, historical*) between *apprentice* and *experienced* collaborators. The third hypothesis, analyzes the editor attributes and implies that experienced editors will edit more articles than others. The third hypothesis leads to the fourth and last hypothesis. It analyzes the fact that experienced editors are more likely to contribute to similar types of articles rather than to dissimilar. Strong correlation is found for editors belonging to the *apprentice* class and for most of the article types. In contrast to our analysis the work by Keegan et al. has as a main focus modeling the network structure of editors and how this reflects on the dynamics of Wikipedia and contemporary and emergent entities and events. On the other hand, in our analysis we focus on larger real world news corpora which inherently represent emerging entities and events. In addition, we also distinguish the lag for different entity types. As a last diverging point in our work, is the analysis of how entities are co-created and its impact on the entity lag.

Entity Interlinking tries to detect links between entities withing a knowledge base. The work done by Nunes et al. [14] uses social network theory measures, such as Katz index to find links between entities. This is related to our work since we analyze the co-referencing of entities within Wikipedia, their collaborator structure and interlinking with events in the Wikipedia's event portal. Such attributes of entities are used to analyze their implications on the entity lag in Wikipedia against news corpora.

First Story Detection typically deals with event onset identification from a stream of text. In [15], Osborne et al., analyze twitter data for first story detection. Wikipedia in this case is used through its entity/event page views to filter tweets that do not represent events. The two sources of information are considered as streams which later on are mapped, by simply checking the spikes of page views for a certain entity/event in a tweet. In our case, the focus is at modeling between two sources of information, Wikipedia and NYT corpus, rather than its usage for story detection.

---

[1] http://en.wikipedia.org/wiki/2011_T%C5%8Dhoku_earthquake_and_tsunami

3. COLLECTION ALIGNMENT

In this section we introduce the experimental setup. To carry out our experiments we first align the two collections, Wikipedia and NYT corpus. The resulting dataset is referred to as the *news-wiki aligned collection* or simply the *aligned collection*. The detailed descriptions of the datasets in our experimental setup are given below:

- **Wikipedia** The *English Wikipedia revision history* [2], whose uncompressed raw data amounts to TBytes, contains the full edit history of the English Wikipedia from January 2001 to December 2013. We consider all versions of encyclopedia articles including versions that were marked as the result of a minor edit (e.g., the correction of spelling errors etc.).

- **News** The *New York Times Annotated corpus* [1] comprises more than 1.8 million articles from the New York Times published between 1987 and 2007. Every article has an associated publication time and we refer to this as the time of the article. Since Wikipedia was released in the year 2001 and our NYT corpora is valid until 2007 we consider sub-collections from both corpora in the time period between the years 2001 and 2007.

- **Taxonomy** The taxonomy from the *YAGO2 Ontology* [7] which combines the clean taxonomy of WordNet with the richness of the Wikipedia category system, assigning the entities to more than 350,000 classes. We also use DBPedia(resource of type `dbpedia-owl:Event`) to collect event pages along with their creation times.

3.1 Preliminaries and Setup

Before delving into detail in the lag analysis, it is necessary to introduce the entity and event notions.

Entity: *We define an entity as something which has a canonical (i.e., uniquely identifiable) representation in Wikipedia. In other words an entity represents a real world concept, e.g.* `People, Organization, Location`*, which might be mentioned in multiple forms in text. We refer to the Wikpedia page dedicated to a given entity as an Entity Page.*

Event: *It is defined as a real-world event that has a Wikipedia article, e.g.* `U.S Elections 2004`*. The Wikipedia article dedicated to the event is referred to as the Event Page.*

We now explain in details the experimental setup. As mentioned before entities are mentioned in text, in this case news, in multiple forms and this sometimes gives rise to the problem of ambiguity, i.e., a given mention potentially refers to more than one entity. One way to resolve such ambiguities, is resolved through the task of *entity linking*. *Entity linking* maps such mentions of ambiguous names onto canonical entities registered in a knowledge base like DBpedia or YAGO. For this task we use *TagMe!* [5].

To maintain high accuracy of the disambiguated entities, we filter out entities with a threshold below 0.3 (values above 0.3 represent high disambiguation scores). Additionally we manually evaluate a random sample of 1000 pairs of disambiguated entities and the corresponding text snippet in the news article. The evaluation took into account whether the disambiguated entity correctly represents the entity in the text snippet. The resulting accuracy of the TagMe! tool after filtering entities, across the different entity types was

on average above 0.9. After filtering the number of disambiguated entities falls to 506,151 from 722,888, with a drop of 30% in the number of entities.

We analyze in total 1.8 million NYT articles, resulting in approximately 506,151 distinct entities. Figure 1 shows the distribution of extracted entities for the years 2001–2007, alongside the number of entities appearing in Wikipedia.
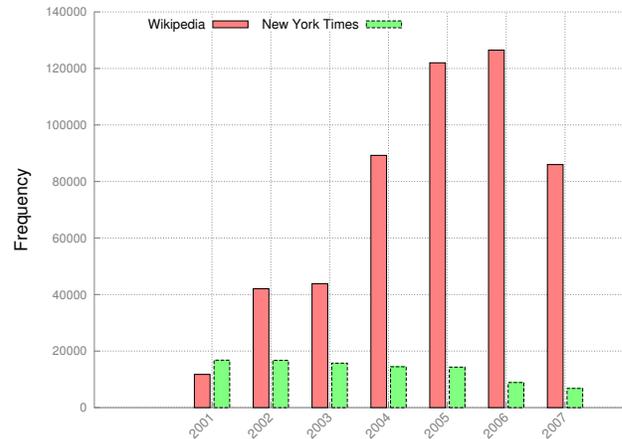

Figure 1: Number of entities appearing in the corresponding years in Wikipedia, and those extracted from the named entity disambiguation process in the NYT corpora.

The final set of entities for our experimental analysis comprises of a collection of 180,478 entities that appear only for the years 2001-2007. Furthermore, the articles are linked to the corresponding state of an entity for the specific year it appears in a NYT articles. For this purpose we make use of the JWPL [19], where given an entity name and a time reference, entity revisions can be retrieved.

4. NEWS REFERENCE DENSITY IN WIKIPEDIA

To start off we want to investigate how news impacts Wikipedia by studying such news references in entity pages. An *entity page* refers to a Wikipedia article dedicated to an entity. Since knowledge bases are reasonable repositories of entities, we compile our set of entities from DBpedia[2]. Entity pages, typically contain references to qualify the stated facts therein. These references are broadly classified into the following sources – `Web, News, Book, Report and Journal, etc.` by Wikipedia[3]. We first study the distribution of news references(of type `News`) in entity pages across multiple *entity categories* and the corresponding entity sections. For this experiment, we first crawled all news articles referenced in the entity pages that are still online. This resulted in a dataset of $129,438$ available news articles out of $411,673$ news references.

News Reference Density: *We define the News Reference Density (NRD) of an entity page, as the fraction of news ref-*
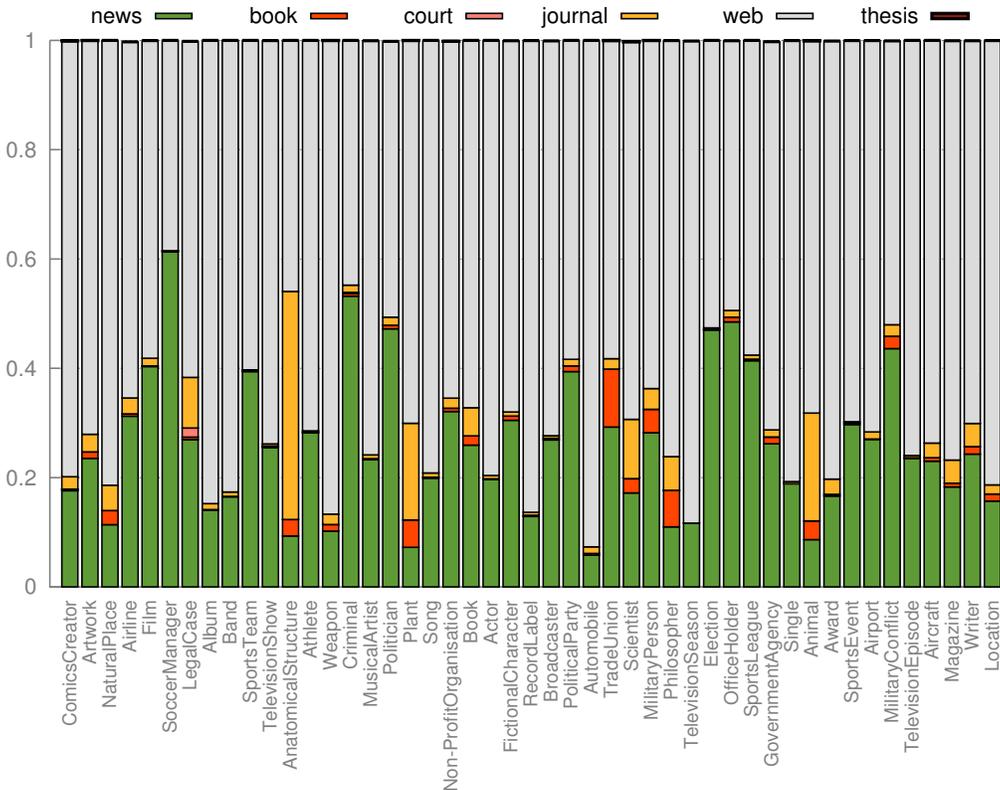
Figure 2: News Reference Density for the different entity categories. The reference density of a given reference type is measured as the fraction of references of that type over all references for the entity page.

erences over all references of all types in the page. Similarly reference densities of other citation types are defined.

We observe that, as expected, most of the references are from the `Web`. However, the second most dominant type of reference are news references constituting 20% of overall references. The NRD varies across entity categories as shown in Figure 2. While the category `OfficeHolders` (mostly politicians) has a high news density, on the other hand `Bands` have high density for web references. The NRD in most cases is stable across years for the different entity categories as shown in Figure 3. However, there are slight variations on the reference density for specialized categories and the corresponding reference types, e.g. category `LegalCase` and `Court` reference types.

Taking into account the organization of Wikipedia entity pages into section, we analyze the distribution of news densities across sections in an Wikipedia entities. We observe that sections in entity pages vary considerably across categories with only some of the sections being common among categories, e.g. '*Early Life*' and '*Career*'. When we look at the partial contribution of the sections to the page news reference density, we observe that while '*Early Life and Career*' in `Politicians` have highest NRD contribution of 64%, the section '*Sports Team*' in `Athletes` has the highest contribution of 19%.

5. ENTITY LAG

For the concept entity we refer to the definition in Section 3. An entity can have multiple ways in which it can be
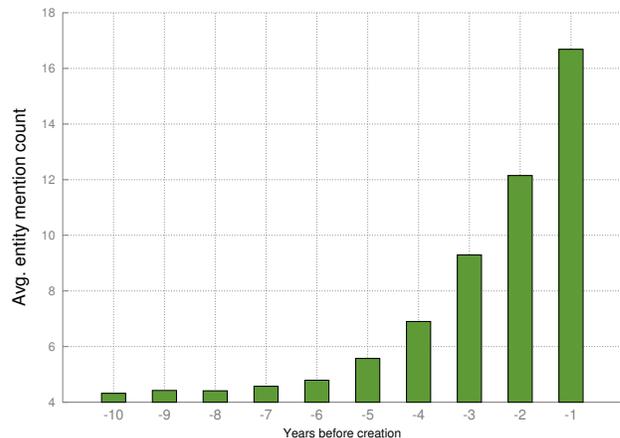

Figure 4: Entity mention counts in news articles before creation of Wikipedia entity page. Mention counts of entities peak a year before it is created in Wikipedia.

mentioned in text. The task of resolving these mentions to the actual entities is a field of *entity disambiguation*, *record linkage* and *entity linking*. We utilize the output of such a linking task to identify entities in our target news corpus and link them to their corresponding entity pages (see Section 3.1).
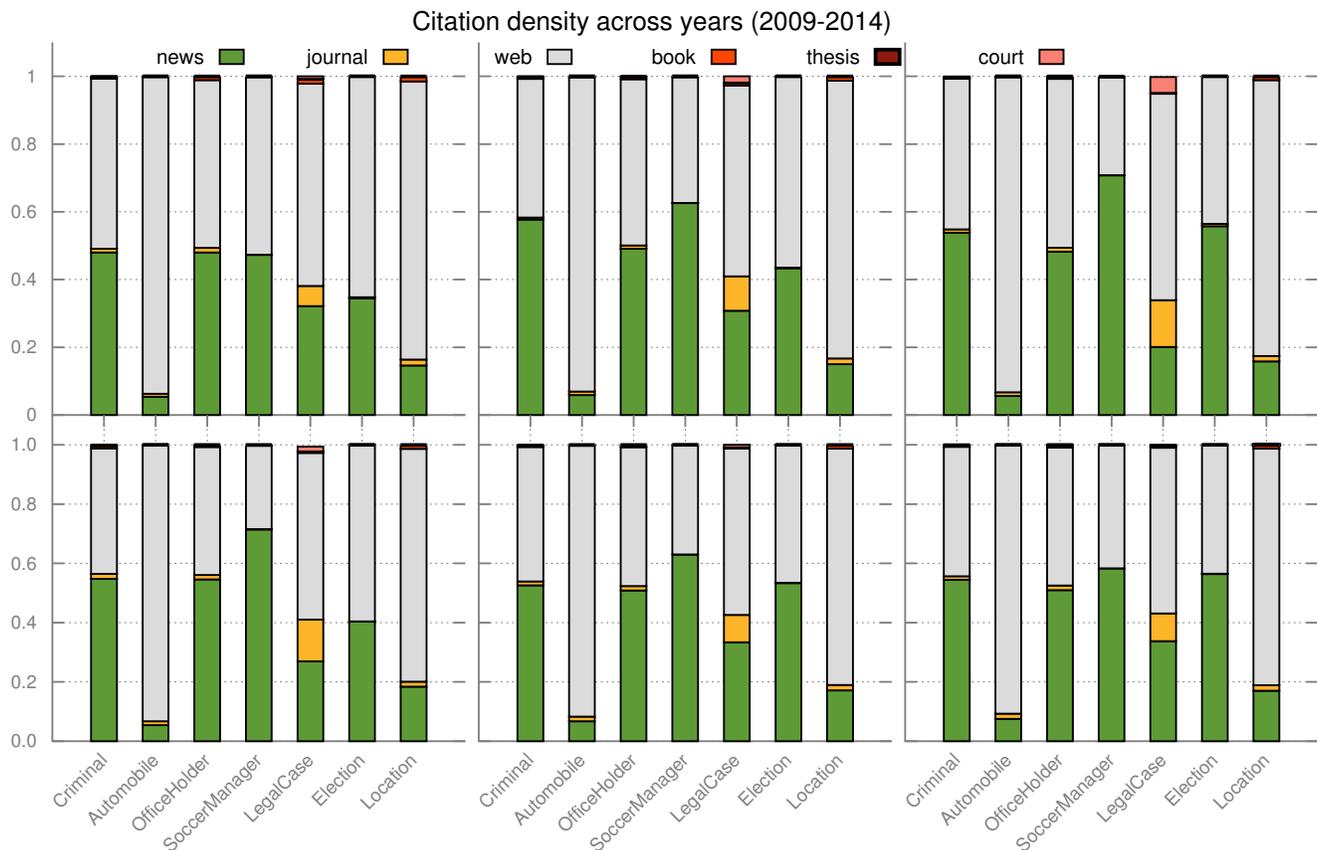
Figure 3: Reference density for the different entity categories. The plots show the reference density for years 2009-2014, in order from left to right.

However, many of the entity pages were created at different points in time. This can be attributed to two factors: *inherent popularity of the entity*, and *evolution of authorship* of entity pages in Wikipedia. One explanation is that entities appearing in authoritative news sources like NYT reflect their popularity. Figure 4 shows the average entity mention distribution (in NYT) across years before the first appearance of an entity in Wikipedia. This follows the assumption that an increase of entity mentions in news sources will eventually result in the creation of an entity in Wikipedia. From Figure 4 it is obvious that shortly before the entity creation in Wikipedia, the entity is mentioned most in news. The second factor, is that Wikipedia's authorship has increased with an ever growing number of editors, hence establishing itself as a independent source of information [11], thus entities can be created from what is deemed as important by the editors in Wikipedia.

To measure the time span between the entity mention and its creation time in Wikipedia we define the *entity lag* below.

Entity Lag: *We define this delay of the first appearance of an entity page relative to the first appearance of the entity mention in a news article as* entity lag *or simply* lag $lag(e_i)$. $\mathbf{lag(e_i) = t_w(e_i) - t_n(e_i)}$*, where $t_w(e_i)$ is the time when the*

first version of entity page of e_i was authored and $t_n(e_i)$ is the publication time of the first mention of e_i in news.

We now proceed to answer the first question of how the creation of entities in Wikipedia lag their mentions in news. We denote the entities which have an absolute lag of less than a month as *low lag entities*, the ones with lag less than a year as *medium-lag entities* and the rest with a lag more than a year as *high-lag entities*. Figure 5 shows the distribution of lag in months for a period of six years.

Second, we see that in the first year of Wikipedia the average lag was high with a majority of entities in Wikipedia lagging behind news. However, quite distinctly, the lag redistributes towards a means of zero in the course of time into a Gaussian or normal distribution. We also see that the absolute number of entities with a lag of zeros go up, and the standard deviation reduces. The lag distribution through the years shifts to a normal distribution, with most of the entities centered around the mean, which in our case is zero. Because Wikipedia only started after 2001, we also consider the entities which were *emergent* in news after 2001 (denoted by the red histogram).

Emerging Entities: *An entity is considered as an* emergent entity *(EE) if its first mention in NYT is after the*
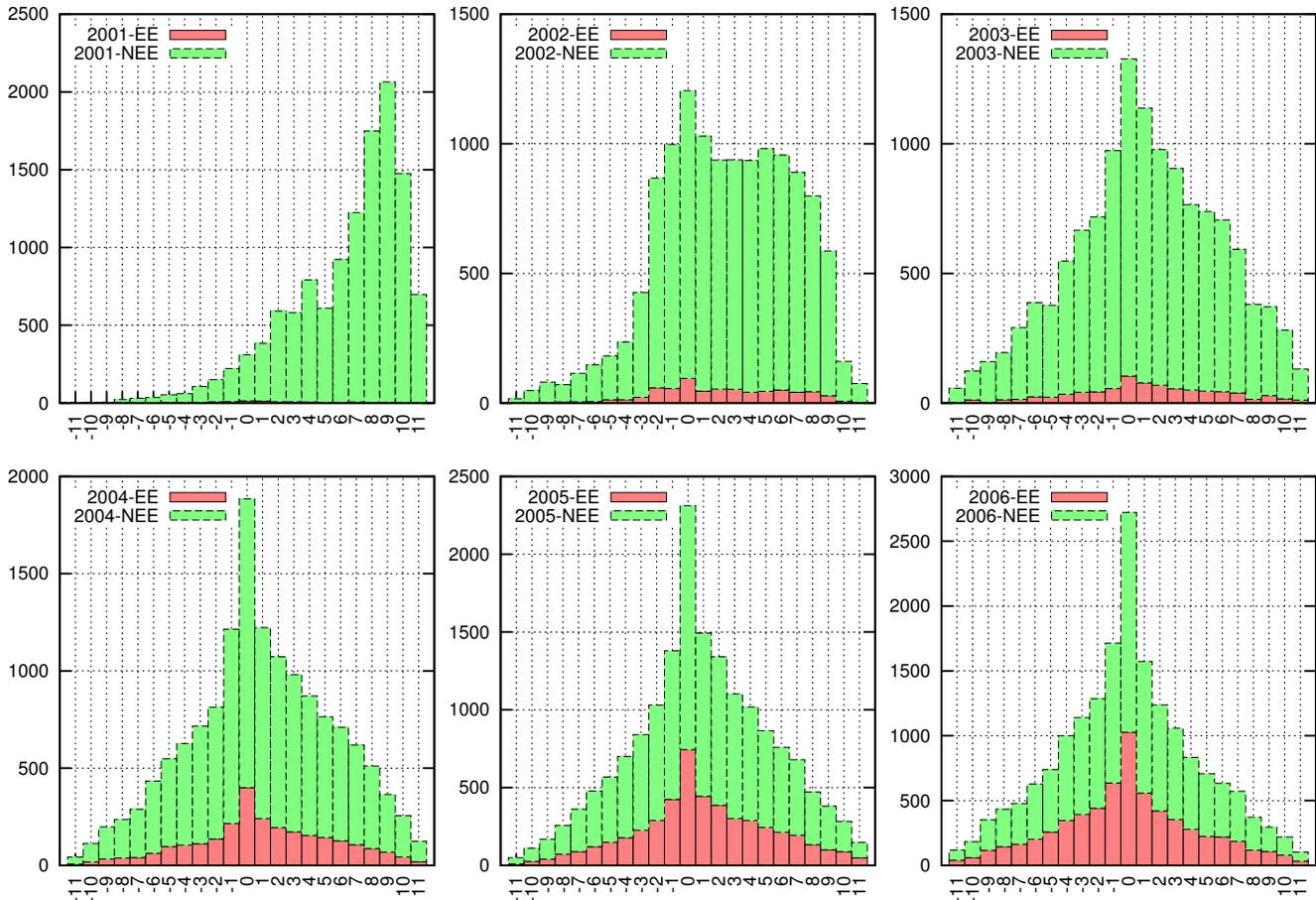
Figure 5: Entity lag in months. The emergent entities are shown in red, they are determined by filtering all entities from the subset of NYT that appear in earlier years before 2001. The y-axis is normalized using the *sum* of entities having medium lag for the emerging and non-emerging entities, respectively.

time when Wikipedia was released, i.e., January 2001.

We observe that the emergent entities, much like the existing entities, have the same distribution. Since news articles are rich in political news and their coverage, we observe that emergent political topics and entities show low lag. An example is `Freedom Fries` which came into prominence in 2003 as a political euphemism for the actual French fries. On the other hand works of fiction like `The lost City` typically exhibit high lag. Similar to the non-emergent entities the lag distribution for *emergent entities* is normal. In Table 1 we test for normality the lag distributions for the non-emergent (NEE) and emergent entities (EE). We use the Shapiro-Wilk test for a *p-value* < 0.05, in which case the hypothesis that the distribution is normal is rejected, otherwise for greater *p-values* the hypothesis is accepted.

Based on the computed distributions, we can already provide a rough estimate of the fraction of *'newsworthy'* entities, which could be missed given a maintenance period. Services that periodically update their entity repositories would lose around half of the entities if their update periods is greater than a month than if they update daily. However, there is not much gain in improving this maintenance period

from 10 months to 9 months.

year	NEE	EE
2001	0.01974	0.00101
2002	0.01305	0.00155
2003	**0.1177***	**0.3585***
2004	0.01127	**0.2196***
2005	0.01009	**0.1091***
2006	0.00269	0.02159

Table 1: Entity lag distribution test for normality. We test whether the distributions come from a normal distribution through the *Shapiro-Wilk test*. The values with * indicate that the lag distribution at the given year is normal.

5.1 Lag for Entity categories

To characterize which entity classes show different lag behavior – positive or negative , low or high – we need to group similar entities which belong to the same semantic category. We attempt to automatically generalize sets of en-

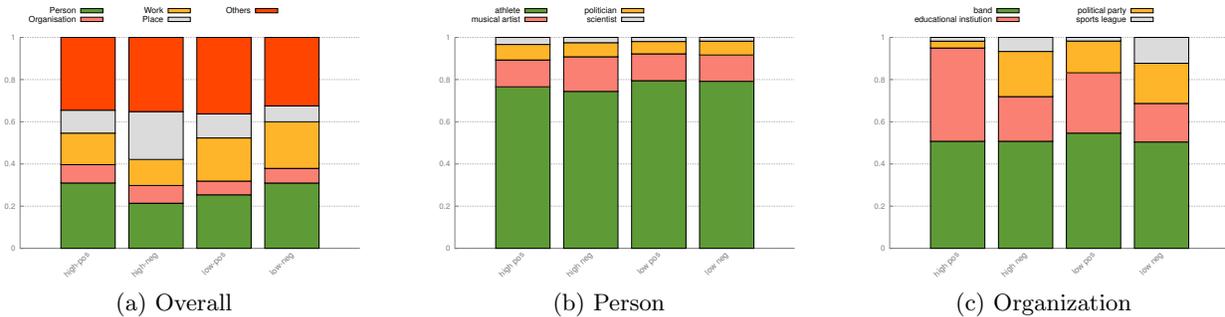**(a) Overall** **(b) Person** **(c) Organization**

Figure 6: Lag distribution of different entity types. The y-axis values are normalized by the sum of the overall entities falling into the different *lag classes*.

tities into meaningful classes based on a pre-existing taxonomy (e.g. YAGO type hierarchy). YAGO infers class memberships from Wikipedia category names, and has integrated this information with the taxonomic backbone of WordNet e.g, `Barack Obama` isA `US President` isA `US Politician` isA `Politician` isA `Leader` isA `Person` isA `entity`.

We first create coarse grained generalizations to obtain the major entity classes. These are *Person, Work, Organization, Places, Other* and are presented in Figure 6(a). High-positive refers to high lag (Wikipedia lags news) whereas high-negative implies a high lead (Wikipedia leads news). It is natural to see that locations (under Places) have the highest negative lag since entity pages for many geographic locations were introduced during the early days of Wikipedia which we refer to as its *bootstrapping period*.

What is interesting is that Wikipedia has a high positive lag for persons (almost 37%) in comparison to other categories. This means that most of the emergent entities are people rather than other entity types. Looking closer into the four major subcategories of people in Figure 8(b), we observe that musicians tend to be mentioned in Wikipedia earlier than news and we confirm that most of them, like the locations, were also created during the bootstrapping period. We then look into the top categories of organizations in Figure 6(c) and make two observations. First, all educational institutions have a high lag and secondly political parties either have a high lead or a small lag. This suggests that political parties are quite popular entities in Wikipedia while educational institutes are not.

The entity class *Work* encompasses all types of books, musical composition and movies. In general *Work* is reported under low lag (around 21%-22%) as compared to its higher lag instances which is around around 12%-14%. In sum, artistic works and locations get reflected in Wikipedia sooner than other categories while Wikipedia lags news for emerging personalities. The overall distribution of entity lag is distributed as shown in Table 2.

lag type	negative (lag)	positive (lead)
high	57.1%	8%
medium	22.2%	11%
low	0.2%	1.1%

Table 2: Absolute entity lag distributions for all lag types. The numbers are aggregated over the years 2001-2006.

6. EVENT LAG

We now turn to studying lag in real world events as documented by news articles. Similar to the definition of entities, we define events as those which have a canonical representation via an *event page* in Wikipedia. Also similar to entities, there might be more than one articles which refer to the same event. We define lag as the publication time difference between the first news article which reports the event and the Wikipedia event page. Events reported in the news can be as a reaction to an event in the past, or a build up to an upcoming event. We do not make a difference in both these cases and treat the first news article reporting the event as the inception of the event.
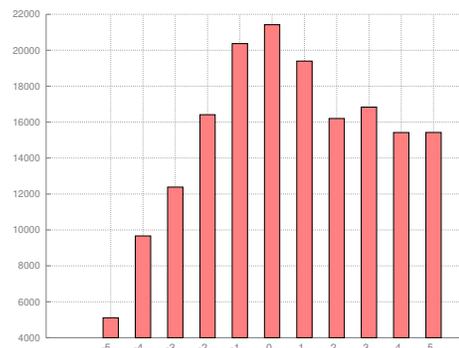

Figure 7: Event news reference lag (in years) in Wikipedia. Most of Wikipedia events fall into *low-lag* class, showing high dynamics of reporting real news events in Wikipedia.

6.1 Emerging Entities in Event Pages

In our final experiment we study how events influence the creation of entity pages in Wikipedia. For this experiment we considered all event pages in DBpedia with their publication time (resource of type `dbpedia-owl:Event`). Unlike the previous experiments we do not rely on the NYT corpora and hence can consider the entire Wikipedia revision history.

The notion of the publication is synonymous to our earlier notion, i.e., the first time the event page was introduced in Wikipedia. We then extracted all the entities in the event page which are explicitly linked (i.e. linked to a valid Wiki-

(a) Emerging Entity Density
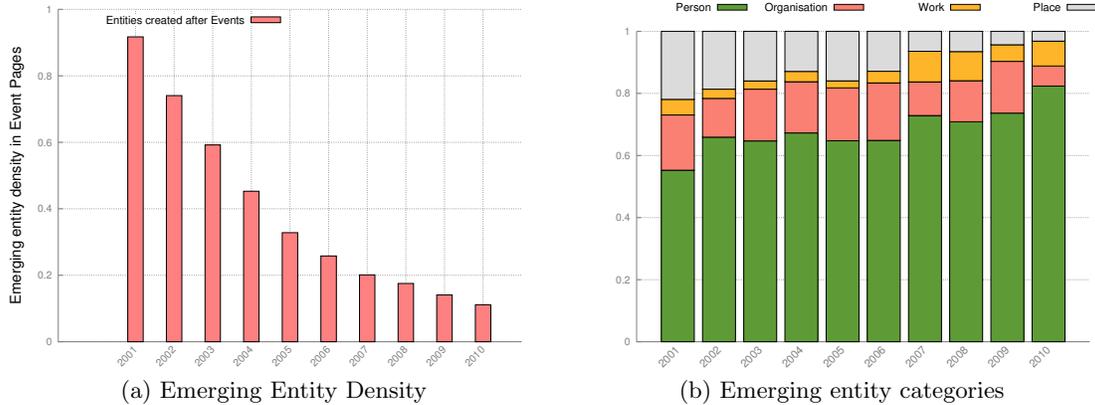
(b) Emerging entity categories

Figure 8: Emerging entity density in Wikipedia event pages.

pedia entity page) in the most current version of the event page. Next, we compared the publication times of the entities mentioned in the events page and the event publication time. To this effect, we make a simplistic assumption about the entities mentioned in the event page: *entities created after the event page are created because of this event.*

Emerging Entities in Event Pages: *Based on this assumption, we define* emerging entity density *of an event page as the fraction of entities which were created after the event page. We refer to such entities as emerging entities (note that this is different from the emergent entities present in the previous section).*

As an example, consider the event page of the "*Charlie Hebdo Shootings*"[4] which was created in 7th January, 2015. The entity "*Corinne Rey*" or "*Coco*" [5] who is mentioned in the event page became popular after the event and subsequently had an entity page created five days later on 12th January.

The emerging entity density (EED) evolution from 2001-2010 is presented in Figure 8(a) where the y-axis represents the average emerging entity density of event pages in a given year. We have a total of 14,604 events with 179,981 entities with the exception of events from the last few years owing to the lack of event data in DBPedia for this period. We see that in the early years the EED of event pages was very high, sometimes above 80%, meaning most of the entities mentioned in the event pages were emerging. Understandably, this declines every year resembling the phenomena of diminishing returns. However, we still see a high percentage of emerging entities in the recent event pages which point to the fact that event pages are great repositories of upcoming and emerging entities missing in the knowledge bases. We also observe that the curve, although decreasing, tends to stabilize in the recent years around 13%. Finally, we look at the categories of emerging entities in Figure 8(b) to find that people comprise the majority of the emergent entities consistently over the years. On the other hand, organizations were emergent between 2001-2005 but their EED contribution to event pages has been decreasing from 2006 onwards.

[4] http://en.wikipedia.org/wiki/Charlie_Hebdo_
shooting
[5] http://en.wikipedia.org/wiki/Coco_(cartoonist)

7. DISCUSSION AND CONCLUSION

Wikipedia is an invaluable resource documenting entities and events and is used as a important input source for constructing knowledge bases. News articles on the other hand, we find are routinely cited in Wikipedia, suggesting that they are high-quality and authoritative sources of facts about entities and events. In this work, we attempt to understand how newsworthy entities and events flow into Wikipedia by defining lag as the inter-appearance time in news and Wikipedia. We use seven years of overlapping news and the Wikipedia revision history to analyze how lag is distributed and how it has evolved over time. We see that the lag distribution is interestingly a normal distribution.

The implications of this study is manifold. First, it shows the promise of news collections as a resource for mining emerging entities. The normal distribution of the entity lag shows that almost 50% of the entities before occurring in Wikipedia are already mentioned in news. Hoffart et. al in [6] have initial attempts for discovering emergent entities in News and Web streams. Secondly, news is an invaluable resource for mining facts about entities and relations between entities. Our experiments on News reference density show that a high proportion of the facts and relations about entities are qualified with a news reference. Additionally our category- and section-wise analysis shows what kind of aspects of which entity-type can be found in news.

Secondly, entity and event repositories relying on Wikipedia can now quantify the degree of loss or re-calibrate their update frequencies based on the lag distribution we provide. Additionally, they can also optimize emergent entity coverage of entities by focusing on event pages. In the earlier years, Wikipedia used to lag more than news in terms of entities, while this slowly converges to a normal distribution over the years. We also observe that the lag for events is far lower than entity pages, which means they get reported far quickly.

Thirdly, we also discover that event pages are nice containers for emergent entities with around 12% of the entities from event pages being emergent pages. There have been studies [17] which have reported the low growth rate in Wikipedia. We attest their finding by showing that event pages in Wikipedia, which contributed to a high number of entities, have a low emerging-entity density in the recent years. However, this low density might be because Wikipedia has

eventually achieved a steady state and yields diminishing returns for new entities.

One of the limitations of the study is that new only consider the New York Times collection which might be biased towards in news coverage and hence in the entity coverage. We hope that, given the size and international nature of NYT, the results might still be representative of the overall effect of news over Wikipedia.

Acknowledgment

8. REFERENCES

[1] New york times annotated corpus. *http://corpus.nytimes.com.*

[2] Wikipedia. *http://en.wikipedia.org/.*

[3] Judit Bar-Ilan and Noa Aharony. Twelve years of wikipedia research. In *Proceedings of the 2014 ACM Conference on Web Science*, WebSci '14, pages 243–244, New York, NY, USA, 2014. ACM.

[4] Christian Bizer, Jens Lehmann, Georgi Kobilarov, SÃűren Auer, Christian Becker, Richard Cyganiak, and Sebastian Hellmann. DBpedia - a crystallization point for the web of data. *Web Semantics: Science, Services and Agents on the World Wide Web*, 7(3), September 2009.

[5] Paolo Ferragina and Ugo Scaiella. Fast and accurate annotation of short texts with wikipedia pages. *IEEE Software*, 29(1):70–75, 2012.

[6] Johannes Hoffart, Yasemin Altun, and Gerhard Weikum. Discovering emerging entities with ambiguous names. In *23rd International World Wide Web Conference, WWW '14, Seoul, Republic of Korea, April 7-11, 2014*, pages 385–396, 2014.

[7] Johannes Hoffart, Fabian M. Suchanek, Klaus Berberich, Edwin Lewis-Kelham, Gerard de Melo, and Gerhard Weikum. Yago2: Exploring and querying world knowledge in time, space, context, and many languages. In *Proceedings of the 20th International Conference Companion on World Wide Web*, WWW '11, pages 229–232, New York, NY, USA, 2011. ACM.

[8] Johannes Hoffart, Mohamed Amir Yosef, Ilaria Bordino, Hagen Fürstenau, Manfred Pinkal, Marc Spaniol, Bilyana Taneva, Stefan Thater, and Gerhard Weikum. Robust disambiguation of named entities in text. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP '11, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[9] Meiqun Hu, Ee-Peng Lim, Aixin Sun, Hady Wirawan Lauw, and Ba-Quy Vuong. Measuring article quality in wikipedia: Models and evaluation. In *Proceedings of the Sixteenth ACM Conference on Conference on Information and Knowledge Management*, CIKM '07, New York, NY, USA, 2007. ACM.

[10] Rianne Kaptein, Pavel Serdyukov, Arjen De Vries, and Jaap Kamps. Entity ranking using wikipedia as a pivot. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, New York, NY, USA, 2010. ACM.

[11] Brian Keegan, Darren Gergle, and Noshir Contractor. Hot off the wiki: Dynamics, practices, and structures in wikipedia's coverage of the tohoku catastrophes. In *Proceedings of the 7th International Symposium on Wikis and Open Collaboration*, WikiSym '11, pages 105–113, New York, NY, USA, 2011. ACM.

[12] Brian Keegan, Darren Gergle, and Noshir Contractor. Do editors or articles drive collaboration?: Multilevel statistical network analysis of wikipedia coauthorship. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work*, CSCW '12, New York, NY, USA, 2012. ACM.

[13] A. Kittur, E. Chi, B.A. Pendleton, B. Suh, and T. Mytkowicz. Power of the few vs. wisdom of the crowd: Wikipedia and the rise of the bourgeoisie. 2007.

[14] Bernardo Pereira Nunes, Stefan Dietze, Marco Antonio Casanova, Ricardo Kawase, Besnik Fetahu, and Wolfgang Nejdl. Combining a co-occurrence-based and a semantic measure for entity linking. In *ESWC*, 2013.

[15] Miles Osborne, Sasa Petrovic, Richard Mccreadie, Craig Macdonald, and Iadh Ounis. Bieber no more: First story detection using twitter and. In *TAIA*, 2011.

[16] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, New York, NY, USA, 2007. ACM.

[17] Bongwon Suh, Gregorio Convertino, Ed H. Chi, and Peter Pirolli. The singularity is not near: Slowing growth of wikipedia. In *Proceedings of the 5th International Symposium on Wikis and Open Collaboration*, WikiSym '09, pages 8:1–8:10, New York, NY, USA, 2009. ACM.

[18] Pu Wang and Carlotta Domeniconi. Building semantic kernels for text classification using wikipedia. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, New York, NY, USA, 2008. ACM.

[19] Torsten Zesch, Christof Müller, and Iryna Gurevych. Extracting lexical semantic knowledge from wikipedia and wiktionary. In *Proceedings of the International Conference on Language Resources and Evaluation, LREC 2008, 26 May - 1 June 2008, Marrakech, Morocco*, 2008.