Information Evolution in Wikipedia

Andrea Ceroni
L3S Research Center
Hannover, Germany
ceroni@l3s.de

Mihai Georgescu
L3S Research Center
Hannover, Germany
georgescu@l3s.de

Ujwal Gadiraju
L3S Research Center
Hannover, Germany
gadiraju@l3s.de

Kaweh Djafari Naini
L3S Research Center
Hannover, Germany
naini@l3s.de

Marco Fisichella
L3S Research Center
Hannover, Germany
fisichella@l3s.de

ABSTRACT

The Web of data is constantly evolving based on the dynamics of its content. Current Web search engine technologies consider static collections and do not factor in explicitly or implicitly available temporal information, that can be leveraged to gain insights into the dynamics of the data. In this paper, we hypothesize that by employing the temporal aspect as the primary means for capturing the evolution of entities, it is possible to provide entity-based accessibility to Web archives. We empirically show that the edit activity on Wikipedia can be exploited to provide evidence of the evolution of Wikipedia pages over time, both in terms of their content and in terms of their temporally defined relationships, classified in literature as events. Finally, we present results from our extensive analysis of a dataset consisting of 31,998 Wikipedia pages describing politicians, and observations from in-depth case studies. Our findings reflect the usefulness of leveraging temporal information in order to study the evolution of entities and breed promising grounds for further research.

Keywords

Entity Evolution, Temporal Information, Events, Wikipedia

1. INTRODUCTION

Exploratory search systems have gained a lot of attention lately, as they help users to search, navigate, and discover new facts and relationships. With the increasing growth of information on the web, exploratory search interfaces are starting to surface. Time clearly plays a central role, but it is still unclear how to leverage temporal information, even if its involvement has been studied in several areas like information extraction, topic-detection, question-answering, query log analysis, and summarization. Time and temporal measurements can help recreating a particular historical period or describing the chronological contexts of a Web archive [3].

Current Web search engine technology, although providing more sophisticated models and algorithms, still assumes a rather static collection, and does not really take the evolution of the Web into account, with content and link structures changing over time, and new versions of pages added continuously [1, 13, 24]. In addition, the evolution of entities and their relationships have an important impact on their retrieval and accessibility. In accordance with retrieval models in general, we need better ways to provide entity-based accessibility to Web archives, considering the evolution of entities and their temporally defined relationships, classified in literature as *events*. Such relationships can be either of predefined type according to an existing schema, or of dynamic type. Conventional approaches fail to detect the latter type, which is common in real world and constitutes the majority of the relationships associated with news events which are often hard to describe, let alone define using a predefined schema [9]. In practice, examples of complex, event-related queries are:

- *When was the first occurrence of Pope Francis in Wikipedia?*
- *How did the German Wikipedia cover the event "outbreak of E.coli in Lower Saxony" during June 2011?*
- *Which were the recent outbreaks of dangerous diseases in Germany?*

Moreover, the identification of the right event granularity and of an adequate event model is not straightforward.

Also, for creating a useful rich semantic layer with event and entity information, it is necessary that events and entities referenced are consolidated between the content objects within a single archive. This does not mean that all information about entities and events has to be consolidated. Rather, which content object refers to the same event or entity should be resolved. This is an important foundation for interlinking information on entities and events extracted from various content objects to form a rich interwoven semantic layer, which can be re-used for navigation or further analysis services.

Although the aforementioned issues can be tracked down in social media platforms such as YouTube, Twitter and Wikipedia, in this paper we will focus on Wikipedia since it covers a large spectrum of real-world events as well as

OpenSym'14, August 27–29 2014, Berlin, Germany.

http://dx.doi.org/10.1145/2641580.2641612.

associated entities generated by a large amount of human-user-generated content. Wikipedia can be viewed as a complex, multi-relational network with different types of temporal links between users and the content, evolving through a variety of interaction mechanisms. Thus, Wikipedia is an interesting collection for temporal information retrieval and entity evolution tracking.

Although the need for focusing research on these issues has been recognized [3], the contributions of this paper provide a complement to aforementioned work in many aspects:

- We propose a *novel time-based* system framework that will significantly advance efficient and effective indexing, retrieval and exploration of temporal information in Wikipedia.
- We present a foundation for a novel class of evolution-aware entity-based enrichment algorithms, and considerably increase the quality of entity accessibility and temporal indexing for Wikipedia.
- We create an Event Repository by combining events from various complementary sources and facilitate its exploration.

The rest of this paper is organized as follows. We discuss related work in Section 2. In Section 3, we propose a system architecture for accessing Wikipedia through events and significant revisions of a wikipage. In Section 4, we show a preliminary analysis motivating our approach. Finally, we draw conclusions in Section 5 and set precedents for future work.

2. RELATED WORK

2.1 Entity Linking and Evolution in Wikipedia

With the diffusion of Wikipedia as a collaborative source of knowledge, a consistent amount of research effort has been spent to automatically build well structured knowledge bases on top of Wikipedia with the aim of linking entity references in texts to the corresponding entity description in the knowledge base. Examples of such trends are observed in YAGO2 [18], DBpedia[1], and WikiTaxonomy [26]. They share the goal of providing high-quality knowledge bases, in terms of coverage, accuracy, and being up-to-date.

In this context, semantics, knowledge, and terminology continuously evolve over time. Moreover, hierarchical relations between entities might change as a consequence of the change of their meaning over time.

YAGO2 addresses the evolution of entities by marking their properties with time information where available. WikiChanges [25], a web-based application designed to plot an article's revision timeline in real time, includes a web browser extension that incorporates activity spark-lines in the real Wikipedia. Authors introduce a revisions summarization task that addresses the need to understand what occurred during a given set of revisions. Also the work in [15] proposes a method for categorizing and presenting edits during a given set of revisions in an intuitive way and with a flexible measure of significance of each individual editor's contributions.

[1] `http://purl.org/ontology/is/inst/dbpedia`

2.2 Event Detection in Wikipedia

Event detection has been applied in many contexts including topic detection and tracking [2], tracking of natural disasters, and event-based epidemic intelligence [14]. In the digital library community, preliminary work [30] has explored the notion of event-based retrieval, returning events instead of documents. This work, while important, is limited in many respects, including 1) the use of a very simple notion of events, 2) evaluation with a small subset of Wikipedia documents, and 3) the use of a simple ranking mechanism such as returning results chronologically. In [28], a system for event discovery and retrieval in multiple data sources, namely, news articles, videos, and micro-blog streams, is proposed. However, the system makes use of a simplified assumption about the time of an event, and employs a clustering approach on a monthly-partitioned sub-collection, with each document linked to just one event.

With a focus on Wikipedia, there has been a large amount of research. A survey, discussed in [23], categorizes and presents the different areas of research to which Wikipedia is relevant. In the earliest work that proposes to exploit the link between Wikipedia and news events [22], the authors notice that after being exposed through press citation, an article gets a lot of traffic and leads users to improve its quality. In [4], the authors probe the hypothesis that pages showing parallel behavior in their edit variance share some similarities, based on the fact that news events trigger edit volume variations in Wikipedia pages.

Analyzing the trends in page view statistics, instead of page edits, the authors of [8] identify concepts with increased popularity for a given time period. By exploiting temporal features such as view counts over time [8] or edit history [16, 17], methods have been proposed for extracting event-related data from Wikipedia.

In a more recent work [11], the process of editing articles is associated with collective memory building from a sociological point of view, studying the evolution of articles pertaining to the North African uprisings.

Recent attempts have tried to extract structured information from different sources in Wikipedia articles (categories and infoboxes [32], free texts [21], etc.), but ignored the temporal dynamics of articles in Wikipedia. A common assumption is a predefined or limited schema of the events detected. In [19], authors demonstrated that articles in Wikipedia about current events exhibit structures and dynamics distinct from those observed among articles about non-breaking events. These findings have implications for how collective intelligence systems can be leveraged to process and make sense of complex information.

3. APPROACH

Wikipedia features a variety of interaction and update mechanisms on the content, as well as different types of explicit and implicit links between users. This leads to constant modifications and updates contributing to various page revisions. In the rest of this paper we will refer to Wikipedia pages as *wikipages*.

Some information generated in a particular time period might no longer be available in a future version of the wikipage entailing the entities involved, due to the edit activity in Wikipedia pages. On the one hand, wikipage revisions are triggered by events in which the corresponding entities are involved. This results in content being either added or mod-

ified over the course of the events. Such activity contributes to the evolution of the wikipage from the content point of view. On the other hand, the entities mentioned in the content of a particular wikipage may also change (evolve) over time. This pertains to both the evolution of related entities (mentioned in the content) and the evolution of the importance of such relevant entities with respect to a wikipage. It is useful to provide users with the possibility of accessing historical information, via a comprehensive evolution-aware view of entities.

In this paper we aim to address entity evolution in Wikipedia by exploiting wikipage revisions. In addition, we investigate the role of entities in *events*. Although it is possible to search the older versions of wikipages, it is a cumbersome endeavour for a user to sift through a number of different versions in order to gauge the evolution of an entity, in terms of the change in content, and even more so with respect to corresponding temporally defined relationships with relevant entities, classified in literature as *events*. We aim to tackle the concomitant challenges by exploiting a temporal index of the wikipages. We believe that providing easy means to temporally query for information associated with entities can also greatly benefit historians, archivists, librarians among other interested parties.

3.1 Architecture

In this section we describe a pipeline of methods that we use to realize our goals. In Figure 1 we portray the architecture of our proposed system. The system brings together two main areas of interest: events and significant revisions of a wikipage. We attempt to offer a comprehensive view over the evolution of entities through the events they were involved in and the significant versions their wikipages went through in their history.

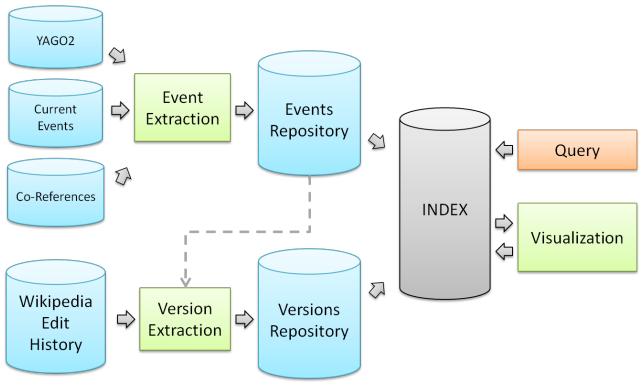

Figure 1: The architecture of our system.

Leveraging information from *YAGO2*, Wikipedia's *Current Events* portal, and *Co-References*[2], the *Event Extraction* component identifies events and introduces them into an *Event Repository*. The *Event Repository* stores events from multiple sources using a unified representation. The events in the repository are indexed and can be retrieved and investigated in the *Visualization* component, either as a result to a *Query*, or as part of an explorative session. Taking into consideration the entire Wikipedia Edit History, the *Version Extraction* component, identifies the *versions* of wikipages (revisions where a significant contribution was made), and stores them in the *Versions Repository*. The versions are in turn indexed and can be retrieved as results to a query or as a part of the interactive explorative visualization. Moreover, the *Events Repository* can be used as a point of reference by the *Version Extraction* component, as significant contributions that alter the wikipage of an entity are highly likely to be made during events that affect the entity. An interface with the end user is achieved via the *Visualization* component, that also can display the results to queries made to our system. In the following sections we describe each of the components in detail.

[2]our method for identifying events, represented as dynamic relationships between entities

3.2 Versions Repository

The creation of a wikipage and the subsequent emergence of numerous revisions through its evolution, result in a vast amount of information regarding the subject of the wikipage. We observe that the number of revisions with respect to various wikipages follows a power law distribution (see Figure 8, later discussed in Section 4.3). This means that a relatively small number of wikipages correspond to a large number of revisions. However, not all the revisions of the wikipage are equally useful. This can be attributed to (i) a large number of vandalism-induced *rollbacks*, and (ii) minor changes resulting in new revisions. The huge volume of revisions on the whole, affect the efficient retrieval of relevant information from corresponding wikipages, and hinder the assessment of their evolution. In order to cope with the vast number of nearly-redundant revisions, that do not add much value in terms of content consumption for a user, we adopt a wikipage versioning scheme. Our versioning approach is presented in the Figure 2.

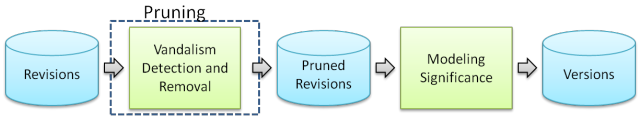

Figure 2: Version Extraction

3.2.1 *Vandalism Detection*

The first step in dealing with a huge volume of revisions corresponding to wikipages, wherein a fair amount of information is redundant in subsequent revisions, is to detect revisions borne out of acts of vandalism. In order to do so, we use a vandalism detection tool for Wikipedia called STiki [34]. STiki helps us to identify vandalised revisions by the following methods.

- Revisions resulting from edits that are undone by an anti-vandalism bot (i.e., ClueBot NG [6]).
- Revisions resulting from edits that are undone by trusted humans using the STiki GUI interface.
- Revisions resulting from rollback actions done by trusted users.

3.2.2 *Significant Revisions*

Having identified vandalised revisions and discarding these from further use, the next step is to identify those revisions

which are significantly different from their preceding and succeeding ones. We define such revisions as *significant revisions*. The notion of significance can be modelled in a number of ways, by considering different indicators.

- The text that is *inserted* or *deleted* leading to new revisions.
- The length of the *modified* (inserted or deleted) text.
- The entities present in the revisions.
- The Cosine Distance between two revisions as a measure of change.
- The presence of particular terms appearing in wikipages.
- Bursts in the edit behavior.
- Proximity to events where the entity was involved
- Methods exploiting the semantic meaning of the contexts in the wikipages.

We can thereby identify the significant revisions of a wikipage. We will further refer to these significant revisions as *versions* of the wikipage.

3.2.3 Periodic Versions

In order to gain an understanding of the evolution of a wikipage, we can also rely on *periodic versions*. Considering a periodic standpoint, with respect to which we can compare different revisions of a wikipage would enable us to view its evolution at a higher level of granularity. For example, we can consider the first revision of a wikipage from each day, each week or month as the periodic standpoint. This is a reasonable way to gauge evolution, since any periodic standpoint can capture evolution, albeit with varying granularities.

3.3 Events Repository

Understanding and identifying the information needs that trigger queries and searches in a document collection is not a trivial task and has been the subject of many works [10, 20, 35]. Web queries and searches are typically classified as *Navigational*, *Informational*, and *Transactional* [5]. However, the categorization of information needs is more complex than that [27], especially for informational queries that are broadly defined as reflecting an intent of acquiring information. Queries belonging to the same category can have different related topics, or can be either time dependent or independent, possibly affecting the performance of the information retrieval system that has to process them.

One remedy to the aforementioned challenges is to adopt a more focused approach, that is optimized with respect to specific kinds of information needs and scenarios. We choose to focus on event-driven searches, where users interact with the system to understand and explore given events. We believe that real-world events affect the information evolution available in an up-to-date and collaboratively edited collection like Wikipedia. In this scenario, exploiting events in tasks like query expansion, document enriching, and ranking would improve the retrieval capability of the system, consequently enhancing the experience of the user in searching for event-related information. The first step in realizing this is creating an event repository that can be used as a starting point.

Building a unique and comprehensive list of real-world events is difficult, if not impossible, since events usually exhibit different complexity (i.e. number of participants, relationships amongst them) and granularity (i.e. "small events" such as the acquisition of a football player by a team are often ignored and not reported in favor of "big events" such as the Arab Spring). We tackle this problem by defining a general event model and exploiting three different and complementary sources of events to populate the Events Repository.

3.3.1 Event Model

Similar to existing event models [29, 33], we describe events in terms of a set of named entities (like people, places, and organizations), a set of terms that describe the event, and a time period. More precisely, an event e in our model is a tuple $e : (entities, description, start, end, source)$, where *entities* is a list of Wikipedia pages participating in the event, *description* is a textual description of the event, *start* and *end* are the starting and ending date of the event respectively, *source* identifies the extraction method used to get the event. We classify entities as people, organizations, artifacts, and locations, since these categories have different meanings and play varying roles in describing an event.

On the one hand, this event model is general enough to be suitable for events extracted from different sources and methods. On the other hand, it captures all the main aspects that characterize events. Our model does not consider event hierarchy, the identification and representation of which, is left for future work.

3.3.2 Event Extraction

As existing event repositories are limited, in terms of number, granularity, complexity, and topic of events, a comprehensive real-word event repository does not exist. In order to go beyond these limitations, we extract events by exploiting different methods and sources, which complement each other to some extent. These are: our event detection algorithm, called *Co-References*, and two event sources, namely Wikipedia Current Events portal[3] and the YAGO2 knowledge base [18].

Co-References. The main idea of the Co-References event detection method is based on the assumption that new edits in a Wikipedia article are indicators of a new event, which is either related to or involves that entity. By following the links in the edits to other Wikipedia pages, we can collect more entities relevant to an event in a particular time period. Thus, we define an event as a *set of entities* if their edits co-refer to one another in a time period σ. The parameter σ represents the time delay until edits of two entities refer to each other. An example for an event extracted by the Co-References algorithm can be the event *Obama;G8;Deauville; in 2011* during the G8 summit in Deauville France. In our preliminary study, we focus on a sample of the politicians dataset (described further in Section 4) in the entire year of 2011. Regarding the time period σ, we choose a 7-day interval since this yielded the most intuitive results in our experiments. In total we detected 242 events from the entities for the politicians dataset of the year 2011.

Wikipedia Current Events. The WikiTimes project[4] provides an index for the Wikpedia's Current Events por-

[3] `http://en.wikipedia.org/wiki/Portal:Current_events`
[4] `http://wikitimes.l3s.de/`

tal, which includes daily summaries of events created by the crowd [31]. WikiTimes contains more than 50,000 events from 2001 − 2013 where each event has a title, name, and description. Furthermore, each event can have a category and a set of entities. Most of the events are from the categories: conflict, politics, and crime, while science and art are less represented. The entities are based on the links inside the event description to the other Wikipedia pages.

YAGO2. The YAGO2 knowledge base [18] is an ontology built from Wikipedia infoboxes and combined with Wordnet and GeoNames to obtain 10 million entities and 120 million facts between them. An example of a fact in YAGO2 is <BobDylan> *wasBornIn* <Duluth>, indicating that the entity Bob Dylan was born in Duluth. When possible, facts are augmented with temporal and spatial information, so that it is possible to know when and/or where a fact took place or occurred. We created an event for each temporal fact according to our event model, treating locations (when available) as entities. Due to the lack of textual descriptions for events, we used the predicate label of a fact as its description: in the previous example, the description would have been "wasBornIn". Moreover, in YAGO2 there are also entities categorized as events themselves, e.g. *2011_Australian_Open*, with further temporal, spatial information as well as participating entities. We created one event for every entity marked as event, where the description is the name of the entity itself and the other event fields are filled by exploiting the available information.

To summarize, *Co-References* detects events by exploiting the edit history of Wikipedia and is able to detect events with different duration, complexity (number of entities and their relationships), and granularity (from a wrestling match to Academy Awards and Egypt Revolution). *YAGO2* contains temporal facts regarding entities (e.g. birth, death, political roles) as well as particular entities categorized as events, resulting in a repository that mostly contains high-level and well-known events, often lacking textual descriptions. On the other hand, *Current Events* portal is contributed by Wikipedia users and contains textual descriptions of daily events: they are reliable, thanks to the high level of control within Wikipedia, and endowed with self explanatory textual descriptions. Note that *Co-References* is an event detection algorithm and it is subject to errors, i.e. not all the detected events are true events. Instead of manually evaluating the output events, which is time expensive, we employ the automatic event evaluation technique described in [7] and retain only those events adjudged as true events.

Given these complementary sources, the merged event repository is broad and more balanced, containing events with different granularity, complexity, and time duration. A given event might be present in the repository as different events coming from different sources, thus providing complementary perspectives of it. This also justifies not performing any duplicate detection on our merged repository: apart from exact duplicate events, any other near duplicate would still provide complementary event-related information.

3.4 Visualization

The visualization component can serve as an interface to show the results of a query, or it can facilitate the interactive exploration of the information stored in the index.

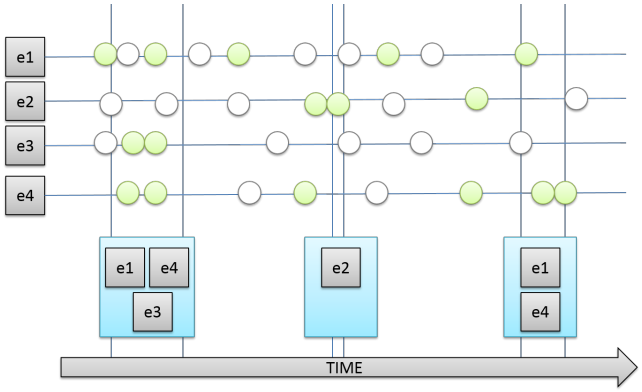

Figure 3: Interplay of events and versions.

In Figure 3 we give an example of the interplay between events and versions. We consider 4 entities whose evolution in time is marked by the appearance of new versions. Each new revision is marked by a dot. If the revision can be considered significant enough to be a version, the dot is colored, otherwise it is white. Moreover, the entities considered are involved in events during their lifetime. For illustration purposes, let us assume that in the studied timespan we encounter 3 events: one event where the participating entities are e_1, e_3, e_4, another one where only e_2 is involved, and finally an event where e_1, e_4 are involved. The figure depicts our assumption that versions are more likely to appear when the entities are involved in events, but this is not the only cause for them. The parallel evolution of wikipages, and their involvement in events is hard to understand, and we aim at facilitating their interpretation. The visualization interface helps the user get insights into how events lead to the evolution of entities, and how this is reflected in the versions. It can also give insights on the reasons for which new versions of the articles appeared, and how this mirrors the actual evolution of the entities.

The interface enables the parallel exploration of the Event and Version Repositories, offering the user a comprehensive view over the evolution of entities, driven by events, and mirrored by the creation of new versions. On the one hand, it enables the user to find all the events where an entity was involved in, and then to identify those versions where the evolution was first reported in. On the other hand, it enables the user to find out if the significant updates that lead to the creation of a new version were related to an event, and if that was the case, it offers an explanation to why and how this was done. Events can be explored in terms of the time period when they occurred, as well as in terms of the entities that participated in it, caused it, or were affected by it. When visualizing an event, the user has the possibility to visualize the versions of the wikipages of its entities as they were at the time of the event, and then understand their evolution as a follow-up to the event. Thus, the deeply interconnected nature of wikipage versions and events becomes more apparent and accessible to the users.

4. ANALYSIS

In this Section, we show an analysis that supports our objectives and approach. First, we describe our dataset, and then we dive into events and entity related analysis.

Source	Total	Politicians	Politicians 2011
All	2,629,740	50,168	1,401
YAGO2	2,578,547	42,399	360
Current Events	50,951	7,527	799
Co-Reference	242	242	242

Table 1: Number of events within the event repository, split by different sources.

4.1 Dataset

We focus our analysis on a set consisting of 31, 998 wikipages registered as entities in YAGO2 [18] and belonging to the *politician* class.

Versioning. In order to identify edits, we made use of the JWPL Wikipedia Revision Toolkit [12]. It solves the problem of storing the entire edit history of Wikipedia by computing and storing differences between two revisions. We used the Wikipedia revision history dump released in October 2012. For every wikipage w, we sample *periodic versions* $w^{(t)}$, on a yearly basis from 2004 to 2012, by accessing the revision history of w. We follow the versioning approach described earlier in Section 3. We first detect and eliminate revisions that are borne out of acts of vandalism. Then we consider periodic versions on a yearly basis,such that the revision on the last day of each year is the version. Note that wikipage revisions are different from the wikipage versions $w^{(t)}$ that we consider. A version $w^{(t)}$ is the wikipage w as it was at time t. Each version $w^{(t)}$ was also annotated with the named entities mentioned in it, extracted through the Stanford CoreNLP parser[5]. An additional *entity* field storing the named entities mentioned in $w^{(t)}$ was created, so that queries could be focused just on entities instead of the full text of $w^{(t)}$. The entire collection of $w^{(t)}$ was indexed with Apache Solr[6] in order to perform further analysis. The index also supports temporal queries and allows to restrict the search to those $w^{(t)}$ matching a desired version date t.

Events. The events extracted as described in Section 3.3.2 have been indexed with Apache Solr. Each event is represented as a document having a field for every element of the event model described in Section 3.3.1 (e.g. description, entities, etc.), allowing to perform queries only on particular subsets of fields. Free-text queries are supported as well, by simply performing queries on the entire content of documents.

4.2 Events

In this Section we analyze the event repository that is constructed as described in Section 3.3, with the goal of showing how the different sources (Co-References, Current Events, YAGO2) contribute to the overall event set and how they differ from each other.

4.2.1 Number of Events

In Table 1 we provide the number of events contained in our repository, along with the contributions from the different sources. As already explained in Section 3.3, the three considered sources cover very different time periods. YAGO2 contains events and temporal facts that occurred even more than 1000 years ago (e.g. Zanzibar <*wasCreatedOnDate*> 1000-01-01), the Current Events method keeps track of events since 2001, and the Co-References method has been restricted to 2011 to get an event set requiring reasonable computational effort while being significant for comparisons. Moreover, the entity sets considered in the sources are different: YAGO2 and Current Events contain almost all the Wikipedia pages, while Co-References has been run on the politician dataset already introduced in Section 4.1. These facts result in very different contributions from each source, as depicted in Table 1. Overall the repository contains more than 2.6 million events, most of them (almost 2.5 million), coming from YAGO2. However, if we restrict the count to those events that occurred in 2011 and involving politicians, then the contributions are more evenly distributed. In order to make a fair analysis, in the following sections we will consider this latter case.

[5] `http://nlp.stanford.edu/software/corenlp.shtml`
[6] `https://lucene.apache.org/solr/`

Source	Active Entities
YAGO2	472
Current Events	310
Co-Reference	366

Table 2: Number of *active* entities within the event repository, split by different sources.

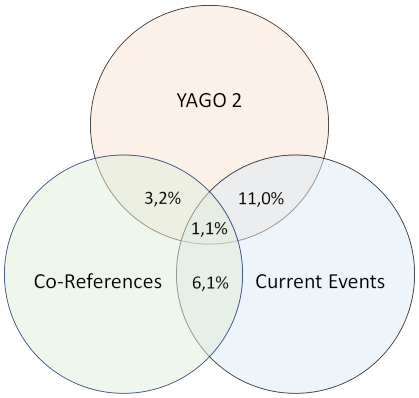

Figure 4: Overlap of the event sources in terms of active entities.

4.2.2 Active and Inactive Entities

A further distinction that we make is between *active* and *inactive* entities: the former participate in at least one event, while the latter do not. Within our politician dataset, composed of 31,998 entities, only 1007 entities are active. This low number can be attributed to at least two reasons. First, we only consider events that occurred in a particular year (2011); there might be politicians that are active in other time periods but not in that year. Second, we noted that our dataset contains many dead politicians, e.g. Vyacheslav Molotov (1890-1986) and Otto von Bismarck (1815-1898), which clearly could not participate in events in 2011. In the rest of our analysis we will consider only active entities.

4.2.3 Entity Overlap

One possible criterion to assess the complementarity between the different event sources is the overlap of their active entities, i.e. how many active entities appear in more than one source. In Table 2 we show the number of active entities contained in each different source. The fact that their sum (1148) is slightly greater than the total number of active

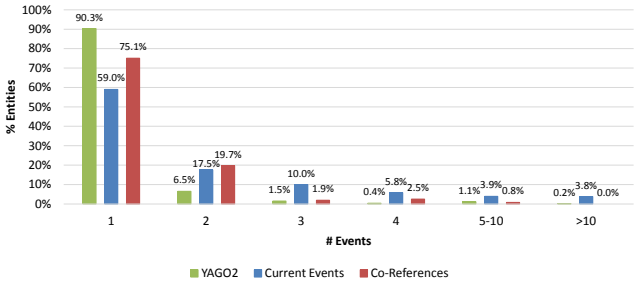

Figure 5: Entity Distribution over number of events they participate to.

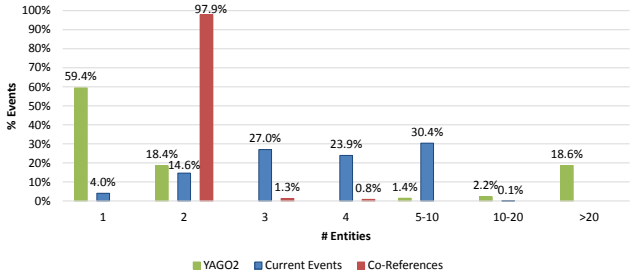

Figure 6: Distribution of events over the number of their participating entities (complexity).

entities (1007) already reveals that the number of active entities that are present in more than one source is quite low. This intuition is confirmed in Figure 4, which shows a Venn diagram for the different event sources. It is possible to observe that the number of overlapping active entities over the three sources represents only 1.1% of their union, showing that the different sources can complement each other. Examples of active entities that are present in all the sources are *Gamal Mubarak*, who was involved in the Arab Spring, and *Vladimir Putin*.

4.2.4 Entity Activity

Active entities can have different *degrees of activity*, based on the number of events they participate in: *strong* entities are those participating in many events, while *weak* entities appear in one or two events. The distribution of entity activity can be a further criterion for describing event sources, since different sources might have entities with different degrees of activity. In Figure 5 we show, for every source, the distribution of entities with respect to the number of events they participate in (i.e. their activity). It is possible to observe that YAGO2 mostly has weak entities (more than 90% of them only participate in 1 event), while Current Events and Co-References contain more dynamic entities. Almost 25% of active entities in Co-References appear in at least 2 events, while for Current Events more than 40% of entities appear in at least 2 events and 7,7% participate in at least 5 events. These facts corroborate the hypothesis that the three event sources considered in our work are complementary.

The most active entity is *Barack Obama*, appearing in 84 events in Current Events, while other examples of dynamic entities are *Silvio Berlusconi* (27 events in Current Events), Vladimir Putin (5 events in Co-References), and *George W. Bush* (5 events in YAGO2).

4.2.5 Event Complexity

Finally, another criterion to describe event sources is the number of entities that their events involve, which we call *event complexity*. Intuitively, non-complex events will involve one or few entities, while more complex events will consist of more entities. Although the complexity of an event is a blurry concept and might be measured in different ways, for instance by considering their temporal duration, we believe that the number of participating entities of an event can be a good indicator of event complexity. In Figure 6 we show, the distribution of events in terms of the number of their participating entities (i.e. their complexity), for each source. Very different patterns can be identified. YAGO2 is highly unbalanced, having almost 60% of events with only one entity and more than 20% with more than 20 entities. The former class corresponds to events representing basic temporal facts like the birth and death of persons, which are composed of only an entity and a date. The latter group represents events that are described by their own Wikipedia page (like *2011 Australian Open* and *Arab Spring*), whose participating entities are those mentioned in the page. Co-References has more than 97% of events with two entities but none with only one entity, since the event detection method is based on finding entities that mention each other in their edits. Current Events is more balanced: more than 95% of its events are quite homogeneously distributed between 2 and 10 participating entities, while very few events consist of only one entity (4%), and more than 10 entities (0.1%). These findings serve as further evidence to how the three event sources are orthogonal and complement each other.

4.3 Edits in Wikipedia

In Figure 7 we plot the number of edits for the wikipages in our dataset on a daily basis, together with the number of wikipages having at least 1 edit during that day. We notice a stable and high number of edits exhibiting a few peaks, after an initial rise. Although the number of edits is high, the edits converge to a small number of wikipages, that seem to plateau around 700 wikipages edited per day. On examining the peaks of edits, we notice that on occasions the number of edited wikipages peak concomitantly. When this does not happen, we get an indication that something significant occurred concerning just some articles, probably triggered by an associated event. The intuition behind this assumption is that the high number of edits was concentrated on a small number of wikipages. Those wikipages accumulate a high number of edits, because they draw more attention than usual. The increased attention of the community is manifested through edits done to the wikipage. The motivation behind this action is probably related to the occurrence of an event in which the entity was involved, that lead to its evolution, most likely an event. To summarize , even the limited set of politicians that compose our dataset exhibits a high number of edits, but they are mostly directed to a small and constant number of wikipages.

In Figure 8 we depict the distribution of all edits with respect to wikipages for our dataset, over the entire period of study. We can notice a power-law distribution. Most wikipages have a small number of edits, or even no edits at all, while only a few wikipages correspond to a large number of edits. There are 470 wikipages over 31,998 exhibiting more than 1,000 edits and 10 of those have over 10,000 ed-

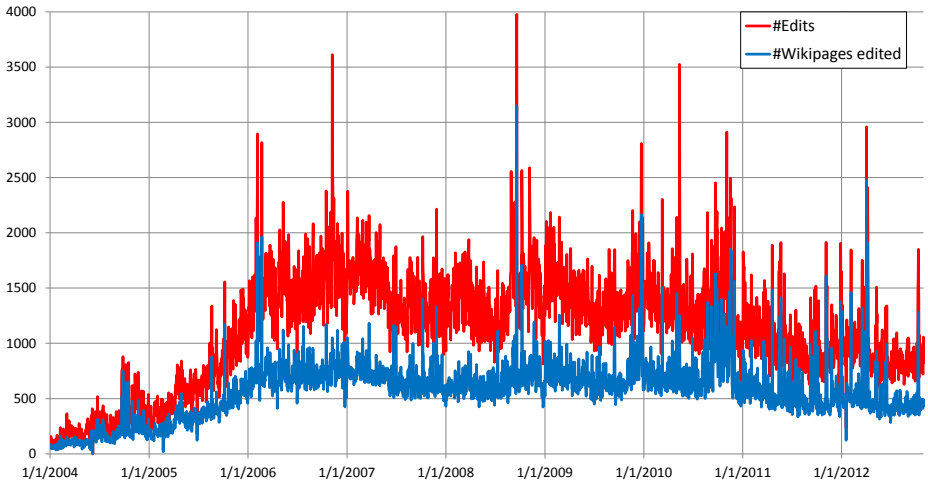

Figure 7: Daily number of edits and wikipages having edits.

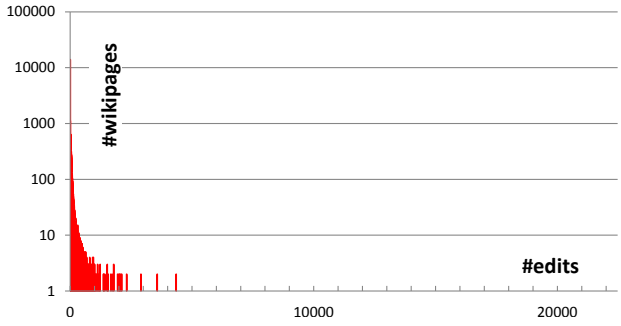

Figure 8: Distribution of #edits.

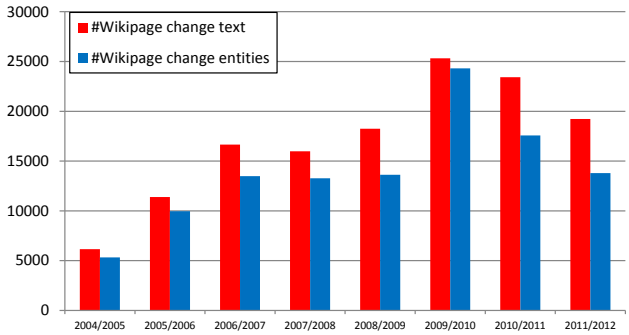

Figure 9: #wikipages having yearly changes.

its. This confirms our observations that the edits are mostly directed to a small number of wikipages, of which the most edits are distributed among an even smaller number of popular wikipages.

We used periodic versions on a daily basis in order to study the nature and dynamics of edits in Wikipedia at a fine granularity, while to capture the evolution of a wikipage over years, we extract yearly *periodic versions* as described in Section 3.

We compare the yearly sampled wikipage versions and investigate how the change evolves in consecutive years. We used Cosine Distance between the text of the versions, as well as the entities contained in it, to quantify change in each case and thereby observe evolution. In Figure 9 we present the average number of wikipages in our dataset that exhibit a change when compared to the version observed in the previous year. Both the change in plain text and in entities depict the same increasing trend. The change in entities is smaller than the change in plain text, showing that wikipages are not just affected by other entities, but also that a large amount of text is dedicated to the description of the actions that those entities comprise. In each consecutive year we can notice an increase when compared to the previous year, that can be explained by the concurrent growth of Wikipedia and its supporting community.

4.4 Temporal Information Retrieval

From an Information Retrieval perspective, wikipage evolution can be perceived by considering how much the top–k document set related to a query changes over time. The more the wikipages evolve over time, the more the retrieved document set should change. Note that changes are computed with respect to wikipedia pages w, not with respect to their revisions $w^{(t)}$. Two top–k result sets are considered to be equal if they contain the same wikipages, regardless of their versions. In order to measure such changes over time, we collected 25 politics–related keywords p (e.g., president, law reform, elections, taxes) and we built a query q for every possible pair (p, t), for instance $q = (taxes, 2005\text{-}01)$ or $q = (taxes, 2007\text{-}01)$. For every query q we performed two different queries q_{text} and q_{entity}, one on the text field

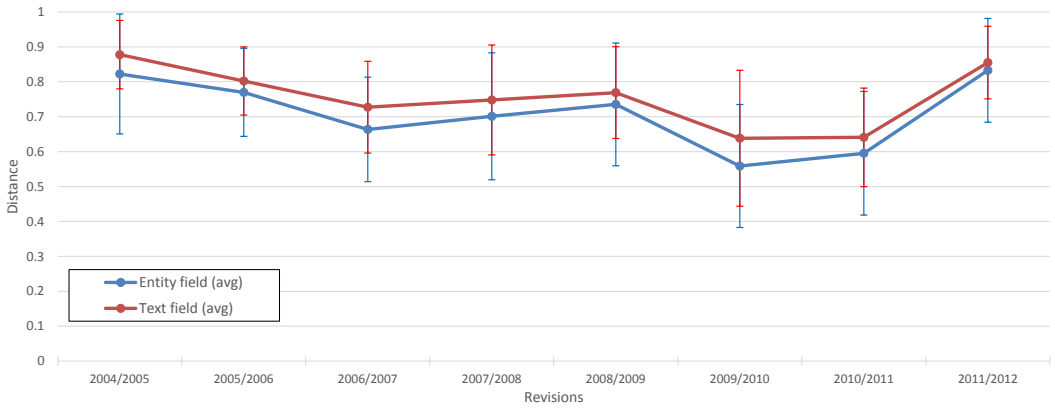

Figure 10: Average differences between consecutive top–10 result sets.

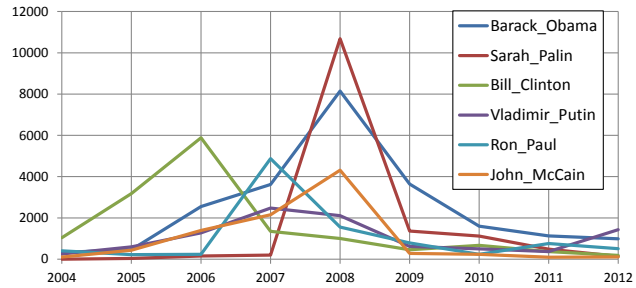

Figure 11: Top wikipages: #edits per year.

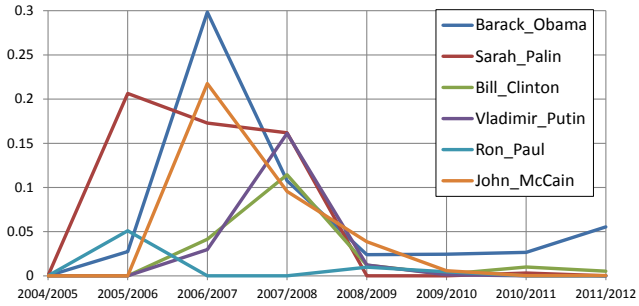

Figure 12: Top wikipages: yearly change.

and one on the entity field of documents $w^{(t)}$. Using tf–idf ranking, we collected the top–10 results for every query and we computed differences between consecutive sets through Jaccard Distance. The results, averaged over the different keywords, are presented in Figure 10. This serves as clear evidence to our premise, that the set of wikipages relevant to a particular query changes over time.

4.5 Edits Volume and Change

From the wikipages that got the highest number of edits over the study period, we choose those corresponding to 6 active politicians: Barack Obama, Sarah Palin, Bill Clinton, Vladimir Putin, Ron Paul and John McCain, and we delve into each of their wikipage evolution. In Figure 11 we plot the number of edits per year for each of the wikipages that we examine. We clearly see that for the 2008 U.S. Presidential Election candidates, Barack Obama and John McCain, there is a burst in the election year 2008. Since our dataset ceases in October 2012, the effects of the 2012 U.S. Presidential Elections are not visible. In Figure 12 we plot the change in text between two consecutive versions for each of our selected wikipages. We notice that the biggest changes occur for the contenders in the U.S. Presidential Elections from 2006 to 2008, as a lot of information becomes available due to the extra attention garnered by the 2 main candidates. Changes continue to occur after the event as a result of summarization in order to fit Wikipedia's encyclopedic style. The highest changes occur between the versions from the end of 2005 to the end of 2008. Between the end of 2008, that was after the elections took place and the end of the subsequent years, the change is minimal.

As most entities exhibit a peak of edits during 2008, the most changes noticed are between the 2007 and 2008 versions. Nevertheless, for Barack Obama and John McCain, the highest changes occur between 2006 and 2007, but these changes come from a smaller number of edits. This shows that looking merely at the number of edits, and observing the peaks is not a strong indicator for significant change in the revisions. The high number of edits might stem from vandalism, minor changes, or edit wars.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce the temporal aspect as a fundamental dimension for capturing entity evolution in Wikipedia. We empirically show that the edit activity on Wikipedia can be exploited to provide evidence of the evolution of Wikipedia pages over time, both in terms of their content and in terms of their temporally defined relationships, classified in literature as events. Furthermore, we present results from our extensive analysis on this dataset and discuss our observations from an in-depth case study.

In the imminent future we will expand the Event Repository by introducing new sources. We will additionally render other versioning techniques briefly described in this paper, and propose novel visualization paradigms.

Our vision is to provide a foundation for a novel class of evolution-aware entity-based enrichment algorithms, and considerably increase the quality of entity accessibility and temporal indexing for Wikipedia.

6. ACKNOWLEDGMENTS

The work was partially funded by the European Commission in the context of the FP7 projects CUbRIK (grant No. 287704), DURAARK (grant No. 600908), the ERC Advanced Grant ALEXANDRIA (grant No. 339233), and by the WikipEvent project.

7. REFERENCES

[1] E. Adar, J. Teevan, S. T. Dumais, and J. L. Elsas. The web changes everything: understanding the dynamics of web content. In *Proceedings of the Second ACM International Conference on Web Search and Data Mining*, WSDM '09, pages 282–291, 2009.

[2] J. Allan, R. Papka, and V. Lavrenko. On-line new event detection and tracking. In *ACM SIGIR*, 1998.

[3] O. Alonso, J. Strötgen, R. Baeza-Yates, and M. Gertz. Temporal Information Retrieval: Challenges and Opportunities. In *International Temporal Web Analytics Workshop,TWAW at WWW*, 2011.

[4] U. Brandes and J. Lerner. Revision and co-revision in Wikipedia. In *Bridging the Gap between Semantic Web and Web 2.0, SemNet*, 2007.

[5] A. Broder. A taxonomy of web search. *SIGIR Forum*, 36(2):3–10, Sept. 2002.

[6] J. Carter. Cluebot and vandalism on wikipedia. 2010.

[7] A. Ceroni and M. Fisichella. Towards an entity-based automatic event validation. In *Proceedings of the 36th European Conference on IR Research*, ECIR, pages 605–611, 2014.

[8] M. Ciglan and K. Nørvåg. WikiPop: personalized event detection system based on Wikipedia page view statistics.

[9] A. Das Sarma, A. Jain, and C. Yu. Dynamic relationship and event discovery. In *WSDM*, 2011.

[10] D. Downey, S. Dumais, D. Liebling, and E. Horvitz. Understanding the relationship between searchers' queries and information goals. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, pages 449–458. ACM, 2008.

[11] M. Ferron and P. Massa. Collective memory building in Wikipedia: the case of north african uprisings. In *WikiSym*, 2011.

[12] O. Ferschke, T. Zesch, and I. Gurevych. Wikipedia revision toolkit: Efficiently accessing wikipedia's edit history. HLT '11, 2011.

[13] D. Fetterly, M. Manasse, M. Najork, and J. Wiener. A large-scale study of the evolution of web pages. In *Proceedings of the 12th international conference on World Wide Web*, WWW '03, pages 669–678, 2003.

[14] M. Fisichella, A. Stewart, K. Denecke, and W. Nejdl. Unsupervised public health event detection for epidemic intelligence. In *Proceedings of the 19th ACM International Conference on Information and Knowledge Management*, CIKM '10, pages 1881–1884. ACM, 2010.

[15] P. K.-F. Fong and R. P. Biuk-Aghai. What did they do?: deriving high-level edit histories in wikis. In P. Ayers and F. Ortega, editors, *Int. Sym. Wikis*. ACM, 2010.

[16] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *ECIR*, 2013.

[17] M. Georgescu, D. D. Pham, N. Kanhabua, S. Zerr, S. Siersdorfer, and W. Nejdl. Temporal summarization of event-related updates in wikipedia. In *Proceedings of the 22nd international conference on World Wide Web companion*, pages 281–284. International World Wide Web Conferences Steering Committee, 2013.

[18] J. Hoffart, F. Suchanek, K. Berberich, and G. Weikum. Yago2: A spatially and temporally enhanced knowledge base from Wikipedia. *Artificial Intelligence Journal, Special Issue on Wikipedia and Semi-Structured Resources*, 2012.

[19] B. Keegan, D. Gergle, and N. Contractor. Staying in the loop: structure and dynamics of wikipedia's breaking news collaborations. In *WikiSym 2012*, 2012.

[20] R. Kumar and A. Tomkins. A characterization of online browsing behavior. In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 561–570, New York, NY, USA, 2010. ACM.

[21] E. Kuzey and G. Weikum. Extraction of temporal facts and events from wikipedia. In *TempWeb 2012*.

[22] A. Lih. Wikipedia as participatory journalism: Reliable sources? metrics for evaluating collaborative media as a news resource. In *the 5th International Symposium on Online Journalism*, 2004.

[23] O. Medelyan, D. Milne, C. Legg, and I. H. Witten. Mining meaning from Wikipedia. *Int. J. Hum.-Comput. Stud.*, 67, 2009.

[24] A. Ntoulas, J. Cho, and C. Olston. What's new on the web?: the evolution of the web from a search engine perspective. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 1–12, 2004.

[25] S. Nunes, C. Ribeiro, and G. David. Wikichanges: exposing wikipedia revision activity. In A. Aguiar and M. Bernstein, editors, *Int. Sym. Wikis*. ACM, 2008.

[26] S. P. Ponzetto and M. Strube. Wikitaxonomy: A large scale knowledge resource. In *ECAI 2008*, Frontiers in Artificial Intelligence and Applications.

[27] D. E. Rose and D. Levinson. Understanding user goals in web search. In *Proceedings of the 13th International Conference on World Wide Web*, WWW '04, pages 13–19, New York, NY, USA, 2004. ACM.

[28] D. Shan, W. X. Zhao, R. Chen, B. Shu, Z. Wang, J. Yao, H. Yan, and X. Li. Eventsearch: a system for event discovery and retrieval on multi-type historical data. In *Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining*, KDD '12, pages 1564–1567, New York, NY, USA, 2012. ACM.

[29] R. Shaw, R. Troncy, and L. Hardman. Lode: Linking open descriptions of events. In A. Gómez-Pérez, Y. Yu, and Y. Ding, editors, *The Semantic Web*, volume 5926 of *Lecture Notes in Computer Science*, pages 153–167. Springer Berlin Heidelberg, 2009.

[30] J. Strötgen and M. Gertz. Event-centric search and exploration in document collections. In *Proceedings of the 12th ACM/IEEE-CS joint conference on Digital Libraries*, JCDL '12, pages 223–232, New York, NY, USA, 2012. ACM.

[31] G. B. Tran and M. Alrifai. Indexing and analyzing wikipedia's current events portal, the daily news summaries by the crowd. *Proceedings of World Wide Web 2014, Web Science Track*, April 2014.

[32] T. A. Tuan, S. Elbassuoni, N. Preda, and G. Weikum. Cate: context-aware timeline for entity illustration. In *WWW 2011*, 2011.

[33] W. R. van Hage, V. Malaisé, R. Segers, L. Hollink, and G. Schreiber. Design and use of the simple event model (sem). *Web Semantics: Science, Services and Agents on the World Wide Web*, 9(2):128 – 136, 2011. Provenance in the Semantic Web.

[34] A. G. West, S. Kannan, and I. Lee. Stiki: an anti-vandalism tool for wikipedia using spatio-temporal analysis of revision metadata. In *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*, page 32. ACM, 2010.

[35] R. W. White, W. Chu, A. Hassan, X. He, Y. Song, and H. Wang. Enhancing personalized search by mining and modeling task behavior. In *Proceedings of the 22Nd International Conference on World Wide Web*, WWW '13, pages 1411–1420. International World Wide Web Conferences Steering Committee, 2013.