# On the Value of Temporal Anchor Texts in Wikipedia

Nattiya Kanhabua and Wolfgang Nejdl
L3S Research Center / Leibniz Universität Hannover, Germany
{kanhabua, nejdl}@L3S.de

## ABSTRACT

Wikipedia has become a widely accepted reference point for information of all kinds; real-world events (e.g., natural disasters, man-made incidents, and political events) as well as specific entities like politicians, celebrities, and entities involved in an event. Due to its open construction and negotiation, Wikipedia is an important new cultural and societal phenomenon, and the content of Wikipedia articles is a valuable source for different applications. For instance, the edit history and view logs of Wikipedia can be leveraged for detecting an event and its associated entities. In this study, we analyze temporal anchor texts extracted from the edit history. We propose a model for Wikipedia and anchor texts viewed as a temporal resource and a probabilistic method for ranking temporal anchor texts. Our preliminary results show that relevant anchor texts composed of evolving information (e.g., the changes of names and semantic roles, as well as evolving context) that reflects societal trends and perceptions, thus being candidates for capturing entity evolution.

## 1. INTRODUCTION

Participative content generation and sharing in Web 2.0 offer new rich data sources for a large scale analysis of patterns in human and especially collective attentions as a crowd phenomenon. The social negotiation and construction processes, e.g., are reflected by early editing activities on pages referring to real-world events [7] as well as by discussions on the talk pages [11]. Previous works exploit edit history [8] and article view logs [3] for detecting events and entities related to the events, which provide promising results towards generating entity-specific news tickers and timelines.

It has been observed that entities, events and concepts are changing over time due to the evolving natures of the social Web and collective attentions [10]. The study of entity evolution comprehends a wide range of IR applications, for example, web archiving, temporal search and longitudinal analytics [1]. For web archiving, the evolution of entities, events, and concepts over time must be addressed already in the gathering, indexing, and retrieval steps to enable their archiving with the desired granularity. Changes must be discovered and quantified to guide these steps. Furthermore, the pace at which the social Web grows requires novel methods that overcome the snapshot paradigm used for archiving traditional web pages and that will shift towards stream processing and summarization.


*TAIA'14,* July 11, 2014, Gold Coast, Australia.


In this paper, we envision that *temporal anchor texts* mined from the edit history of Wikipedia can be used as a hook for tracking entity evolution. The reason for this is twofold. First, anchor texts are useful complementary description for target pages, which are widely applied to improve Web search results [4, 5]. In more detail, anchor texts can be regarded as a short summary (i.e., a few words) of the target document, which represents *collective wisdom* from people other than the author of the target document and captures aboutness or what the document is about. We believe that adding the time dimension in anchor text mining can possibly help keep up with the temporal development of entities, events and concepts. Second, exploiting only anchor texts, rather than the whole document content, enables a large-scale analysis using limited resources, and a more robust discovery of emerging contexts and evolving information.

There are three main research questions that are interesting for the task of tracking and detecting entity and event evolution, namely, (1) What evolving information can be tracked, e.g., the changes of names and semantic roles, or emerging new context?, (2) How to automatically aggregate and generate a timeline describing the conceptual evolution of entities?, and (3) How to track the evolution of long tail entities over time? In this study, we propose to analyze temporal anchor texts in Wikipedia as a starting point for knowing what aspects of entity evolution can be observed from collective wisdom, and to what extent such information available.

Our contributions are: 1) defining *entity evolution* as the task of tracking and detecting relations between mentioning of the same entity in Wikipedia over time, 2) a model of Wikipedia and anchor texts viewed as a temporal resource, 3) a probabilistic method for ranking temporal anchor texts in Wikipedia, and 4) preliminary results showing that temporal anchor texts are useful for entity evolution detection.

## 2. ENTITY EVOLUTION

**Problem Statement.** Over time, entities and events gradually evolve, which comprise both the evolution of the entities and events themselves and the evolution of their representations. In general, such evolutions can reflect in two main issues: *terminology change* and *context change*. Terminology evolution refers to the changes of words related to their definitions, semantics, and names of persons, places or organizations. It is important to note that terminology evolution is a continuous process caused by two major problems: 1) spelling variation in the modern and historic language [6], and 2) semantics or concepts drift over time, i.e., new words are introduced, others disappears, or word sense and meanings change [12]. Common changes in context would include

changes in organizational roles, personal relationships, and world-related knowledge, e.g., geo-political changes. Moreover, we can anticipate changes occurring in different periods of life or in workplace and societal settings.

**Tasks/Challenges.** It is important to adequately deal with such evolutions, for example, to *track* and *detect* changing properties of entities and events over time. Therefore, we envision tasks related to entity evolution and their expected outcomes including (1) correcting and extending the representation and models of entities and events in a self-contained fashion coping with evolution in semantics, use contexts, and interpretations, and (2) developing methods for tracking semantic and terminological shift over time as well as detecting major changes in knowledge structures that are relevant for a change in interpretation or meanings of entities and events.

A key challenge to tackle this problem includes a careful identification of selected context dimensions, which information or features (e.g., properties of the entities) should be captured and affected by the respective evolution. This process of selecting adequate features is expected to create varying results depending upon the type of information and the use case under consideration. To this end, it is expected to be able to identify some core time travel representations depending on types of evolution supported by the designed and implemented methods and tools to deal with information evolution relevant to observe evolutions and adapt context information.

# 3. EVOLUTION MODEL

In order to track and detect evolution, we propose to exploit collective intelligence, i.e., anchor texts in Wikipedia, for easing human appraisal. We tackle this problem by extending a temporal model for Wikipedia and anchor texts originally proposed in [9]. The problem of automatically tracking and detecting the evolution of entities and events can be split into two different sub problems. First we need to identify and represent the relation between entities and their properties and intended context at a given time. We call such a representation: *entity snapshot*, which is always based on a given document collection. Second, we need to perform a fusion of different entity snapshots by identifying the relations between their evolving properties and concepts.

In our context, a document collection is Wikipedia $\mathcal{W}$ that consists of a set of articles or pages, $\mathcal{P} = \{p_1, \ldots, p_n\}$. We categorize Wikipedia pages into three main types: 1) named entity pages describing a concept about people, organizations, or countries, 2) event pages containing information about past, ongoing or anticipated events, and 3) those that do not describe a named entity or an event, e.g., user talk pages, category pages, etc. We call a page in the first type *a named entity page*, and the term "entity" and "named entity" are used interchangeably for simplicity. An entity $e_i$ is represented by terms constituting the title of an entity page. Each page $p_i \in \mathcal{P}$ consists of: 1) terms $\{w_1, \ldots, w_n\}$ , and 2) a time interval $[t_a, t_b]$, i.e., a time period that $p_i$ exists in the collection: $p_i = \{\{w_1, \ldots, w_n\}, [t_a, t_b]\}$.

A page $p_i$ is composed of a set of its revisions $\{r_j | r_j \in \mathcal{R}_i\}$. A revision $r_j$ consists of: 1) a set of terms $\{w_1, \ldots, w_m\}$, and 2) a time interval $[t_c, t_d)$ or $\mathit{TInterval}(r_j)$ associated to $r_j$: $r_j = \{\{w_1, \ldots, w_m\}, [t_c, t_d)\}$. Note that a time interval of any $r_j$ excludes its last time point, $[t_c, t_d) = [t_c, t_d] - \{t_d\}$. By partitioning $\mathcal{W}$ with respect to a time granularity $g$, we will have a set of Wikipedia snapshots $\mathbb{W} = \{W_{t_1}, \ldots, W_{t_z}\}$. In this work, we use a one-month granularity.

Each Wikipedia snapshot $W_{t_k}$ consists of the current revision $r_c$ of every page $p_i$ at time $t_k$, such that, $W_{t_k} = \{r_c | \forall p_i : r_c \in \mathcal{R}_i \wedge t_k \in \mathit{TInterval}(r_c) \wedge \cap \mathit{TInterval}(r_c) \neq \emptyset\}$. Let $\mathcal{A}$ be a set of anchor texts of all entities in $\mathcal{W}$, $\mathcal{A} = \{a_1, \ldots, a_m\}$. We define $\xi_{i,j}$ as a relationship between an entity $e_i$ and its anchor text $a_j$, that is, $\xi_{i,j} = (e_i, a_j)$. Each entity-anchor relationship $\xi_{i,j}$ has an associated time interval $[t_\alpha, t_\beta]$, i.e., a time period that $a_j$ is added or mentioned as an anchor text of $e_i$. We define $A_{t_k}$ as *an anchor snapshot* as a set of entity-anchor relationships at a particular time $t = t_k$. Thus, $A_{t_k} = \{\xi_{1,1}, \ldots, \xi_{n,m}\}$ where $t_k \in \mathit{TInterval}(\xi_{i,j})$.

# 4. OUR APPROACH

This section presents our approach to identifying named entity articles and extracting entity-anchor relationships. Finally, we propose a time-dependent, probabilistic method for ranking anchor texts.

## 4.1 Temporal Anchor Text Extraction

For each Wikipedia snapshot $W_{t_k}$, we identify all entities in $W_{t_k}$. A result from this step will be a set of entities $E_{t_k}$ at a particular time $t_k$. After that, we determine a set of anchors for each identified entity providing a set of entity-anchor relationships or an anchor snapshot $A_{t_k} = \{\xi_{1,1}, \ldots, \xi_{n,m}\}$.

In the first step, we are only interested in an entity page that can be identified using the approach of Bunescu and Paşca [2]. Given a snapshot $W_{t_k}$ and a set of pages existing at time $t_k$, $W_{t_k} = \{p_i | \forall p_i : t_k \in \mathit{TInterval}(p_i)\}$, the recognition of an entity $p_e$ is based on these heuristics:

- If multi-word title with all words capitalized, except prepositions, determiners, conjunctions, relative pronouns or negations, consider it an entity.
- If the title is a single word, with multiple capital letters, consider it an entity.
- If at least 75% of the occurrences of the title in the article text itself are capitalized, consider it an entity.

After identifying a set of named entities $E_{t_k}$, as the next step we will extract anchor texts from article links for each entity page in $E_{t_k}$. More precisely, for a page $p_i \in W_{t_k}$, we list all internal links in $p_i$ but only those links that point to an entity page $p_e \in E_{t_k}$ are interesting. We then obtain a set of entity-anchor relationships. By accumulating a set of entity-anchor relationships from every page $p_i \in W_{t_k}$, we will have a set of entity-anchor relationships at time $t_k$.

Both steps are processed for every snapshot $W_{t_k} \in \mathbb{W}$. Finally, we will obtain a set of entity-anchor relationships for all snapshots: $\mathbb{A} = \{A_{t_1}, \ldots, A_{t_z}\}$.

## 4.2 Temporal Anchor Weighting

After extracting anchor texts for each snapshot, the weights of aggregated anchor texts will be computed and used for ranking them by importance with respect to a target entity. For a given entity $e$, the set of temporal anchor texts $A_{e,t}$ contains all the unique anchor texts of $e$'s incoming links at time $t$. Each anchor text $a$ is associated to a weighting function $f(a, e, t)$, which can be calculated as the count of inlink pages to $e$ or using a better estimation taking into account the relationship among sites or domains, i.e., considering inlink pages from the same site/domain as *internal links* and those from different domains as *external links* [4, 5]. In this work, we only use anchor texts extracted from within English

Wikipedia articles, thus ignoring site/domain relationships. Consequently, we propose our weighting method that is link-independent and based on a probabilistic model presented in [5]. The link-independent model assumes that the inlink pages to a target page are independent, and they are equally important to the target page. Thus, our weighting function $f(a, e, t)$ can be computed as follows.

$$f(a, e, t) = \alpha \cdot P(a, e, t) \propto P(a, t) \cdot P(e|a, t) \quad (1)$$

where $\alpha$ is a multiplier used to integrate values of $f(a, e, t)$, which can be ignored in ranking because it gives the same value for any anchor and entity pair. The prior probability of an anchor text $P(a, t)$ is treated as uniform across all anchor text at any time snapshot and thus it will be ignored in this paper. For future work, we will investigate how to estimate a prior probability to different anchor texts, e.g., consider the prior probability $P(a, t-1)$ at the previous snapshot. The probability $P(e|a, t)$ will be computed based on the whole collection of Wikipedia entity pages at time $t$ in two manners using: 1) article links, and 2) distinct inlink pages.

$$P_{link}(e|a, t) = \frac{\sum_{a \in A_{e,t}} freq(a, e, t)}{\sum_{a' \in A_t, e' \in E_t} freq(a', e', t)} \quad (2)$$

where $freq(a, e, t)$ is the raw frequency or count of *links* pointing to a destination page or an entity article $e$ with anchor text $a$. Then, the probability $P_{link}(e|a, t)$ is a fraction of link counts in the collection. Similarly, $P_{page}(e|a, t)$ can be computed as the percentage of the raw frequency or count of *distinct inlink pages* pointing to a destination page or an entity article $e$ with anchor text $a$.

$$P_{page}(e|a, t) = \frac{\sum_{a \in A_{e,t}} page(a, e, t)}{\sum_{a' \in A_t, e' \in E_t} page(a', e', t)} \quad (3)$$

In addition to determining weights at different snapshots, it is interesting to predict the trend of an anchor text or its future popularity using a prediction method as presented in [4]. However, we leave this aspect for a future study.

## 5. EXPERIMENT

In this section, we describe our experimental settings and discuss our analysis results.

**Experimental Settings.** Our dataset was created using a dump of English Wikipedia edit history from the Internet Archive[1]. This dump consists of pages and revisions in XML format created between 03/2001 to 03/2008 with the total decompressed size of 2.8 Terabytes. We partitioned the edit history into snapshots using a one-month granularity resulting in 85 snapshots (03/2001, 04/2001,..., 03/2008). In addition, we obtained four more dumps from specific time snapshots, namely, 05/2008, 07/2008, 10/2008, and 03/2009 resulting in 89 snapshots in total. For processing the dumps, we used an open-source library MWDumper[2] to extract all pages, revisions and temporal anchor texts from the dump file, and stored them in MySQL databases.

**Analysis Results.** The final dataset of entities and distinct anchor texts comprises 473,829 and 1,503,142 respectively, where we found on average 3.17 anchors per entity with the maximum number of anchor texts per entity is 564.

[1]http://www.archive.org/details/enwiki-20080103
[2]http://www.mediawiki.org/wiki/Mwdumper

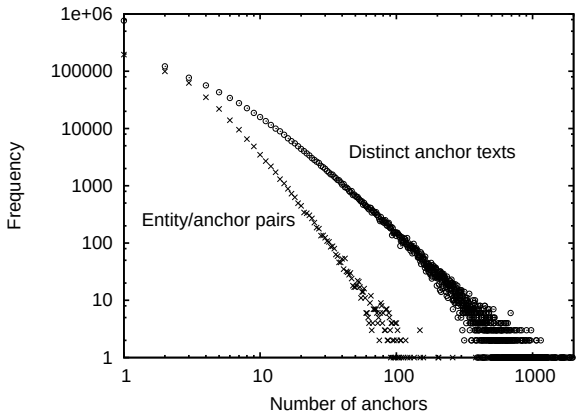


**Figure 1: Distribution of distinct anchor texts, and entity-anchor relationships.**

Figure 1 shows the resulting distributions of distinct anchor texts, as well as entities and their corresponding anchor texts for the latest snapshot 03/2009. As expected, both distributions are long-tail power law distributions.

In order to investigate what constitute entity and event evolution, we considered the top-100 list of entities ranked by the number of their corresponding anchor texts (from the snapshot 03/2009). Then, we manually selected important entities (i.e., politicians, celebrities, companies, geographic names, and technologies) and high-impact events, and ignored the rest of entities due to the limitation of page space in this paper. Figure 2 depicted the selected entities, where the actual ranks are provided in parentheses. For example, the entities September 11, 2001 attacks and George W. Bush are the 2nd and 7th ranked entities with the number of anchor texts of 383 and 150, respectively. We believe that, in large part, the entities with many associated anchors can provide more insights about their evolving properties.

In order to understand this assumption, we determined the similarity of a title (or entity) and an anchor text using Jaccard coefficient because of its simplicity and high feasibility for a large scale collection, such as, the Wikipedia history. The more value of Jaccard coefficient, the higher the similarity between title and anchor text. As illustrated in Figure 2, in general results shows that the more number of associated anchor texts, the less similarity between the title and anchor texts. In this case, the trend of similarity scores (estimated using an exponential regression) increases slightly when the number of anchor texts decreases.

Table 1 depicts evolving information represented by anchor texts and corresponding time snapshots for selected entities including Barack Obama, Hillary Rodham Clinton, Bangalore, Burma, Prince (musician), Pope John Paul II, Pope Benedict XVI, IPod, PlayStation, and Microsoft Windows. Note that, we present the anchor texts ranked by the link-based method because of its better performance compared to the page-based method observed in manually annotating. We removed anchor texts nearly duplicate to the entity terms, which are determined using edit distance with the specified threshold of 3 in our experiment. For example, the anchor texts *Barack Obama* and *Barack Obama's* are very similar to the entity name and they will be removed from the ranked list. Moreover, anchor texts that lexically overlapping with an entity name in their whole terms will be also removed, e.g., the anchor texts *Barack* or *Obama* are excluded from the list.

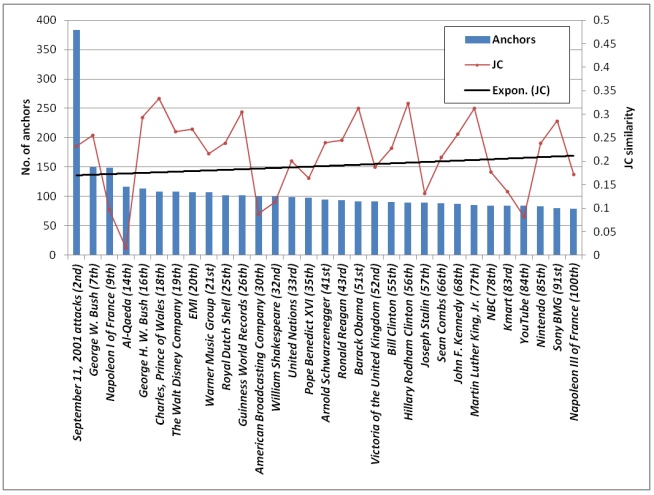


**Figure 2: List of top entities ranked by the number of their corresponding anchor texts.**

In the case of Barack Obama, we observe the changing of his role from a senator to the President. The changing of political roles can be observed also for other entities, such as, Hillary Rodham Clinton, Angela Merkel, and George W. Bush. The collective view about Pope Benedict XVI consists of the name *Cardinal Joseph Ratzinger* used before his papacy, whereas the views about Pope John Paul II reflect different aspects associated to the papacy. The evolving information of geo-location entities that can be detected using our method are geographic name changes for entities like Bangalore, Burma, and Cambodia. In the aspect of changing names over time, we also find temporal anchor texts are good representations for entities in the category of celebrity and entertainment, namely, Prince (musician), Cat Stevens, Sean Combs, and Chad Johnson. In addition, there are many anchor texts referring to trends, e.g., popular songs or current shows/performance of artists. For IT products or software, evolving information is more related to new products or software versions. Examples of such entities are IPod, Memory Stick or Microsoft Windows.

To summarize our findings, temporal anchor texts can be used for capturing entity evolution to some extent, such as, the changing of names or roles, as well as evolving context (e.g., related products or events).

## 6. CONCLUSIONS AND FUTURE WORK

In this paper, we analyzed temporal anchor texts extracted from the edit history of Wikipedia. As shown by experiment results, temporal anchor texts can represent a hook for tracing entity evolution. However, there is a clear need for further study, e.g., a distinction between contemporaneous and historical events/entities. Moreover, it is crucially important to not just identify the facts that are being added, but what time point they are actually anchored. Our plan for future work includes: 1) adding semantic information to anchor texts, 2) improving the changing time periods by considering time mentions surrounding anchor texts, and 3) analyzing event-related articles to discover evolving information for a specific event.

**Acknowledgments** This work was partially funded by the European Commission for the FP7 project ForgetIT and the ERC Advanced Grant ALEXANDRIA under the grant numbers 600826 and 339233, respectively.

**Table 1: Examples of entities with evolving information captured by temporal anchor texts.**

| Entity | Time | Anchor Text | Weight |
|---|---|---|---|
| Barack Obama | 200807 | Senator Barack Obama | 2.71E-07 |
| Barack Obama | 200807 | Illinois Senator Barack Obama | 3.39E-08 |
| Barack Obama | 200807 | U.S. Sen. Barack Obama | 1.69E-08 |
| Barack Obama | 200903 | President Barack Obama | 7.59E-07 |
| Barack Obama | 200903 | President Obama | 7.27E-07 |
| Barack Obama | 200903 | Senator Barack Obama | 9.69E-08 |
| Hillary R. Clinton | 200807 | Senator Hillary Clinton | 1.52E-07 |
| Hillary R. Clinton | 200807 | New York Senator Hillary Clinton | 3.39E-08 |
| Hillary R. Clinton | 200903 | Senator Hillary R. Clinton | 6.46E-08 |
| Hillary R. Clinton | 200903 | Secretary Hillary Clinton | 1.62E-08 |
| Hillary R. Clinton | 200903 | Secretary of State Hillary Clinton | 1.62E-08 |
| Bangalore | 200611 | Bangalore East (Indiranagar) | 1.57E-07 |
| Bangalore | 200612 | Bengaluru | 1.50E-07 |
| Burma | 200504 | Burmese | 2.19E-05 |
| Burma | 200505 | Myanmar | 6.19E-07 |
| Burma | 200505 | Nation of Burma | 6.19E-07 |
| Prince (musician) | 200602 | Prince and the Revolution | 2.64E-07 |
| Prince (musician) | 200604 | mysterious man in purple | 2.20E-07 |
| Prince (musician) | 200608 | Jamie Starr | 1.81E-07 |
| Prince (musician) | 200609 | Prince: Alter Ego | 3.44E-07 |
| Prince (musician) | 200609 | Artist | 3.44E-07 |
| Prince (musician) | 200609 | New Power Generation | 1.72E-07 |
| Pope John Paul II | 200402 | Ioannes Paulus PP. II | 8.85E-06 |
| Pope John Paul II | 200402 | Karol Cardinal Wojtyla | 2.95E-06 |
| Pope John Paul II | 200402 | Karol Wojtyla | 2.95E-06 |
| Pope John Paul II | 200504 | Pope John Paul II's health | 6.84E-07 |
| Pope John Paul II | 200507 | John Paul the Great | 5.00E-07 |
| Pope John Paul II | 200507 | predecessor's | 5.00E-07 |
| Pope Benedict XVI | 200506 | Joseph Cardinal Ratzinger | 1.03E-05 |
| Pope Benedict XVI | 200506 | the new pope | 5.70E-07 |
| Pope Benedict XVI | 200506 | Joseph Alois Ratzinger | 5.70E-07 |
| IPod | 200510 | iPod/iPod Mini | 3.84E-07 |
| IPod | 200606 | video iPod | 2.03E-07 |
| PlayStation | 200202 | Sony PlayStation | 4.11E-04 |
| PlayStation | 200309 | PS | 5.23E-06 |
| PlayStation | 200408 | PSX | 1.19E-06 |
| PlayStation | 200601 | PS1 | 6.55E-06 |
| PlayStation | 200603 | PS2 | 2.42E-07 |
| PlayStation | 200903 | PS3 | 1.62E-08 |
| Microsoft Windows | 200210 | Windows CE | 1.60E-05 |
| Microsoft Windows | 200212 | Windows XP | 9.44E-06 |
| Microsoft Windows | 200601 | Windows XP Pro | 2.85E-07 |
| Microsoft Windows | 200601 | Windows 2000, XP, and Server 2003 | 2.85E-07 |
| Microsoft Windows | 200607 | Windows 98/Me, 2000, XP, Vista | 1.94E-07 |
| Microsoft Windows | 200612 | Windows 9x, ME, NT4, 2000, XP, 2003, Vista and x64 versions | 1.50E-07 |

## 7. REFERENCES

[1] O. Alonso, J. Strötgen, R. A. Baeza-Yates, and M. Gertz. Temporal information retrieval: Challenges and opportunities. In *Proceedings of the First Temporal Web Analytics Workshop*, 2011.

[2] R. C. Bunescu and M. Pasca. Using encyclopedic knowledge for named entity disambiguation. In *Proceedings of EACL '06*, 2006.

[3] M. Ciglan and K. Nørvåg. WikiPop: personalized event detection system based on Wikipedia page view statistics. In *Proceedings of CIKM '10*, 2010.

[4] N. Dai and B. Davison. Mining anchor text trends for retrieval. In *Proceedings of ECIR '10*, 2010.

[5] Z. Dou, R. Song, J.-Y. Nie, and J.-R. Wen. Using anchor texts with their hyperlink structure for web search. In *Proceedings of SIGIR '09*, 2009.

[6] A. Ernst-Gerlach and N. Fuhr. Generating search term variants for text collections with historic spellings. In *Proceedings of ECIR '06*, 2006.

[7] M. Ferron and P. Massa. Psychological processes underlying wikipedia representations of natural and manmade disasters. In *Proceedings of WikiSym '12*, 2012.

[8] M. Georgescu, N. Kanhabua, D. Krause, W. Nejdl, and S. Siersdorfer. Extracting event-related information from article updates in wikipedia. In *Proceedings of ECIR '13*, 2013.

[9] N. Kanhabua and K. Nørvåg. Exploiting time-based synonyms in searching document archives. In *Proceedings of JCDL '10*, 2010.

[10] A. Mazeika, T. Tylenda, and G. Weikum. Entity timelines: Visual analytics and named entity evolution. In *Proceedings of CIKM '11*, 2011.

[11] C. Pentzold. Fixing the floating gap: The online encyclopaedia wikipedia as a global memory place. *Memory Studies*, 2(2):255–272, 2009.

[12] N. Tahmasebi, G. Gossen, N. Kanhabua, H. Holzmann, and T. Risse. NEER: An Unsupervised Method for Named Entity Evolution Recognition. In *Proceedings COLING '12*, 2012.