A bag-of-entities approach to document focus time estimation

Christian Morbidoni and Alessandro Cucchiarelli

Dipartimento di Ingegneria dell'Informazione, Università Politecnica delle Marche, Ancona, Italy
c.morbidoni@univpm.it, a.cucchiarelli@univpm.it

Abstract. Detecting the document focus time, defined as the time the content of a document refers to, is an important task to support temporal information retrieval systems. In this paper we propose a novel approach to focus time estimation based on a bag-of-entity representation. In particular, we are interested in understanding if and to what extent existing open data sources can be leveraged to achieve focus time estimation. We leverage state of the art Named Entity Extraction tools and exploit links to Wikipedia and DBpedia to derive temporal information relevant to entities, namely years and intervals of years. We then estimate focus time as the point in time that is more relevant to the entity set associated to a document. Our method does not rely on explicit temporal expressions in the documents, so it is therefore applicable to a general context. We tested our methodology on two datasets of historical events and evaluated it against a state of the art approach, measuring improvement in average estimation error.

Keywords: focus time; temporal mining; information retrieval; bag-of-entities; linked data; wikipedia; dbpedia

1 Introduction

The growing interest in exploiting the temporal dimension of text documents to improve information retrieval tasks led to a relatively new field of research, referred to as Temporal Information Retrieval [2]. Given that most web search queries express implicit or explicit temporal needs, a considerable amount of research in the field has been made to best answer temporal queries, addressing both recency-sensitive queries, where the users need is to get fresh information, and time-sensitive queries, where the information need is related to a particular point or period in time. Characterizing documents under a temporal dimension is therefore important to support temporal aware search engines and spans two distinct aspects: document creation time and document focus time. While considerable effort has been made in literature to address the first one, few works investigated document focus time estimation [2].

The focus time of a document is defined as the point in time (instant-based focus time), or time interval (interval-based focus time), to which the content

of the document refers and is related to the meaning and the semantics of its content. The concept was introduced in [10] where a general methodology to estimate focus time of text documents is presented. The approach, extended and further evaluated in [11], attempts to detect the focus time of a document by deriving words-time association from a large training corpus of on-line news.

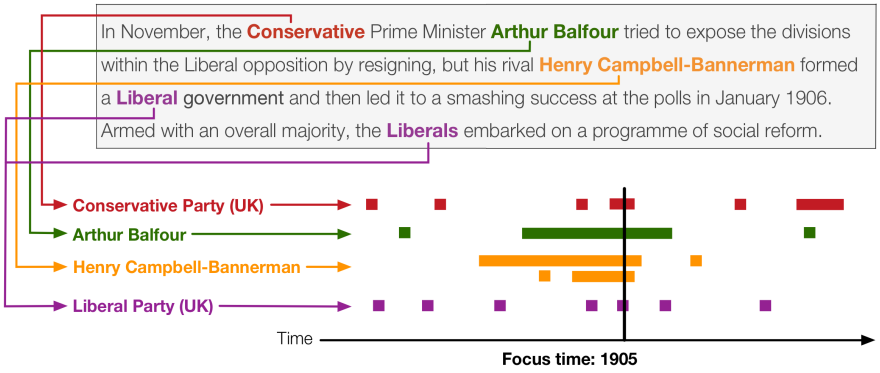

Fig. 1. An example of entities in a document and their relation to time.

To introduce our approach let us consider the short text document in Figure 1 as an example. The document mentions a number of named entities and can be concisely represented as a bag-of-entities. Each entity has associations in time with years and intervals of years, which are derived from the entities representations available on the web. Our assumption is that the focus time of a document is likely to be the point in time relevant to the biggest subset of entities representing the document.

In this paper we start from this simple intuition and rely on a bag-of-entities representation to investigate if and how one can leverage openly available textual and machine readable data to extract meaningful entities-time associations and ultimately develop a fast, unsupervised and reliable method to estimate document focus time. Our method does not make use of explicit temporal expressions in documents, but rather aims at determining the focus time based on the semantics of the content. We think that leveraging named entities and Linked Data has the clear advantage of avoiding possibly expensive computation to learn word-time associations. Entity-time associations are in fact derived by simply parsing wikipedia articles and DBpedia resources.

In our experiments we used a state of the art NERD tool[1] [17] and defined a set of pragmatic rules to extract and rank relevant dates and time intervals associated to each detected entity. We then combined these associations to rank dates and estimate the focus time with a granularity of one year.

[1] https://dandelion.eu/semantic-text/entity-extraction-demo/

We evaluate the approach on two datasets extracted from history related books and web sites, and by comparing our results with the method proposed in [11] (B1) and with a simple method based on explicit temporal expressions in documents (B2). We found that our approach outperforms B1 in average estimation error and considerably increases recall when compared to B2, since in real world dataset explicit dates are not always present.

This paper is organized as follows. In section 2 we discuss related works. In section 3 we describe the different phases of our method. Then we describe our experiments in section 4 and provide an evaluation of the approach on test data in section 5. We conclude the paper and mention possible improvements in section 6.

2 Related Works

According to [2], a document has mainly two temporal dimensions: its timestamp, or creation time, and its focus time. While the first one has been largely addressed in literature, e.g., in [12], few works address the second. Related works, as [19], face the problem of identifying the most relevant temporal expression in a document. However, relying only on temporal expressions is not a good strategy as they can be rare in the text or even poorly related to the content itself.

The first work addressing focus time estimation without relying on temporal expressions, which inspired this research, is [11]. Authors introduced the notion of focus time in [10] and extensively evaluated a generic approach to estimate the focus time of a document based on weighted associations between terms in the document and years. Such associations are statistically extracted from a large news corpus by analyzing sentences that contain explicit temporal expressions and weighting associations based on the co-occurrences of words and dates in sentences. In addition, words are weighted with respect to their discriminative capabilities - using temporal entropy and temporal kurtosis measures - as well as their relevance, using different techniques, including TextRank [15]. Finally the focus time of a document is calculated as the date that maximizes the sum of associations with all the words in the document.

As the approach is based on a learning phase, one possible limitation is that it needs a training corpus which properly covers the event occurring in the documents to be estimated. This means the domain has to be known. For example authors restrict their domain of interest to historical events related to 5 countries in the time range 1900-2013. In this paper we investigate an alternative approach based on a bag-of-entities representation of documents, where each named entity is linked to Wikipedia and DBpedia. Entities-time associations are then derived by parsing a relatively small number of documents (i.e. the Wikipedia and DBpedia entries for all the entities in a document). In our experiments we aim at understanding whether we can combine state of the art named entity recognition with knowledge from Wikipedia and DBpedia to

estimate focus time with reasonable precision, without the need for a learning phase and in domain independent manner.

In [14], the authors attempt to temporally classify images on the web by analyzing the text surrounding them. The work is related to ours in that they extract named entities from text. However, they use the dates associated to such entities (extracted from the YAGO knowledge base [8]) only to filter explicit temporal expressions found in the text, removing those not associated with a mentioned entity.

Other works related to this research come from the temporal information retrieval (TIR) area. [1] and [3] introduce the notion of temporal clusters, which are related to our paper as they can be possibly used to associate time spans to documents. In [20], authors proposed an approach to identify the most relevant temporal expressions in a text document. The relevances of the temporal expressions, detected using the HeidelTime temporal tagger, are calculated with a set of predefined heuristics based on document and corpus-based features. This task is related to ours in that identifying relevant time expressions can give clues to detect the focus time of the document. In general, however, explicit temporal expressions are not always present.

A number of research made use of bag of entities representations to address tasks such as document classification [21] and clustering [9]. In the latter work, a concept thesaurus based on the semantic relations (synonym, hypernym, and associative relation) extracted from Wikipedia is leveraged to enhance traditional content similarity measure for text clustering. A bag of concepts is also used in [7] to improve performances of text classification tasks in a biomedical domain. A bag-of-entities representation, the one used in our methodology, has been used to improve semantic search engines in [4] and to drive recommendation systems in [13], where advantages of this representation in reducing complexity are highlighted. Named entities extracted from documents were used to detect temporal patterns of collective attention in online news, twitter and Wikipedia page views[18].

To the best of our knowledge no research has been conducted on the feasibility of using bag-of-entities and available online open data (specifically Wikipedia and DBpedia) to address document focus time estimation.

3 Methodology

3.1 Singling out the bag-of-entities for a document

The first step in our approach is to derive a bag-of-entities to represent the document. To do so we processed each document via the Dandelion API[2] [17], obtaining a set of spots in the text with links to Wikipedia and DBpedia, along with a confidence score. Dandelion is an evolution of TAGME [6] and, albeit initially designed for short texts (e.g. micro-blogs), it was proved to be effective also for longer texts [5] [16].

[2] https://dandelion.eu/docs/api/datatxt/nex/v1/

The obtained entities are identified by the corresponding Wikipedia page URL, from which it is straightforward to derive the URL of the corresponding DBpedia resource. For each entity we collected the plain text of the corresponding article using the Wikipedia API[3], and we retrieved the RDF representation from DBpedia.

3.2 Enriching entities with temporal data

For each entity we want to retrieve the associated exact dates (years) and time intervals, to then score them with respect to their relevance. We use a date granularity of one year.

We first extract years mentions from the Wikipedia full article and from the short entity abstract contained in DBpedia[4]. The DBpedia abstract is extracted from the Wikipedia article and represents a concise description/summary of the entity. As the dates in the abstract are generally included in the full Wikipedia article, we in fact boost dates that appear in the abstract, as we assume they are in general more relevant. To extract years mentions we tried two options: i) using SUTIME temporal tagger[5] and ii) using a simple regular expression[6]. We observed very few differences in performances of our method, the second option providing slightly better results. For brevity's sake, in this paper we only show results obtained using our simple regular expression.

As an additional source for relevant years we exploit the DBpedia Linked Data representation by spotting triples carrying temporal information. We extracted such triples by querying the DBpedia SPARQL endpoint to get all the properties used in the dataset that have an rdfs:range[7] explicitly declared of type date (xsd:date, xsd:dateTime, xsd:gYear, etc.[8]). We counted the occurrence of each property in the DBpedia graph and selected the ones used at least 10 times, obtaining a set of 113 properties. For each entity in the document, the values of such properties are retrieved and years mentions are distilled and added to the candidates.

If single years intuitively represent dates of short or specific events involving the entity, time intervals represent their periods of activity or existence (e.g. life of a person, rise and fall of a government, etc.). We attempt to derive time periods from the Linked Data (LD) representation of the entity, by considering couples of properties that identify time intervals. For example, for a resource of type Person (the most frequent in the datasets), DBpedia represents his/her period of existence as follows:

```
dbpedia:Audrey_Hepburn dbo:birthDate "1929-05-04".
dbpedia:Audrey_Hepburn dbo:deathDate "1993-01-20".
```

[3] https://en.wikipedia.org/w/api.php
[4] the value of the *dbpedia* : *abstract* property
[5] http://nlp.stanford.edu/software/sutime.shtml
[6] We used the following regex: *[1-2][0-9][0-9][0-9]*
[7] the range of a property is the type of values that it can assume
[8] XML datatypes, https://www.w3.org/TR/xmlschema11-2/

For our experiments, we manually selected from the list of temporal properties 14 couples indicating time ranges and which are used in at least 1000 triples. These include, for example, *birthDate/deathDate* for persons, and *foundingYear/dissolutionYear* for organizations.

3.3 Ranking entities-dates associations

As a next step we rank the relations among the entities and the candidate dates by combining different contributions:

- The occurrences of a date in the full Wikipedia article corresponding to an entity;
- The occurrences of a date in the entity abstract
- The occurrences of a date in the RDF temporal triples of an entity LD representation
- The matching of a date with the relevant time periods for an entity.

We proceed as follows. For each document d in our dataset we define $NE_d = \{e_1, e_2, ..., e_n\}$ as the set of NEs in the document, $w(e_i)$ as the Wikipedia article associated to each e_i and $TW_{e_i} = \{t_1, t_2, ..., t_h\}$ as the set of dates found in $w(e_i)$. Then we calculate the relevance of a date t_j with respect to $w(e_i)$ as:

$$W_{rel}(t_j, e_i) = \frac{freq(t_j, w(e_i))}{\sum_{k=1}^{h} freq(t_k, w(e_i))} \tag{1}$$

where $freq(t_j, w(e_i))$ is the number of occurrences of t_j in $w(e_i)$.

In the same way, we define $TDP_{e_i} = \{t_1, t_2, ..., t_g\}$ as the set of dates found in the e_i entity abstract $dp(e_i)$, and the relevance of a date t_j with respect to $dp(e_i)$ as:

$$DP_{rel}(t_j, e_i) = \frac{freq(t_j, dp(e_i))}{\sum_{k=1}^{g} freq(t_k, dp(e_i))} \tag{2}$$

where $freq(t_j, dp(e_i))$ is the number of occurrences of t_j in $dp(e_i)$. Likewise if $TDT_{e_i} = \{t_1, t_2, ..., t_f\}$ is the set of dates extracted from the RDF triples $tr(e_i)$ related to e_i, the relevance of a term with respect to the data in RDF triples is:

$$DT_{rel}(t_j, e_i) = \frac{freq(t_j, tr(e_i))}{\sum_{k=1}^{f} freq(t_k, tr(e_i))} \tag{3}$$

We consider as set of candidate focus times for the document the set $T_{cand} = TW_{e_i} \cup TDP_{e_i} \cup TDT_{e_i}$. The next step is scoring the elements in T_{cand} looking for matches in the set of time periods extracted from DBpedia, that we call

$TPDP_{e_i} = \{tp_1, tp_2, ..., tp_l\}$. We thus define the relevance of a date in T_{cand} with respect to the entity time-periods in $TPDP_{e_i}$ related to e_i as:

$$TP_{rel}(t_j, e_i) = \frac{\sum_{k=1}^{l} in(t_j, tp_k)}{l} \quad (4)$$

where $in(t_j, tp_k)$ is equal to 1 *IFF* the date t_j is included in the time interval tp_k and 0 otherwise, and l is the dimension (number of elements) of the set T_{cand}. Now we can combine (1), (2), (3) and (4) to define the cumulative measure of relevance of the date t_j with respect to the entity e_i:

$$Rel(t_j, e_i) = \alpha W_{rel}(t_j, e_i) + \beta DP_{rel}(t_j, e_i) + \gamma DT_{rel}(t_j, e_i) + \delta TP_{rel}(t_j, e_i) \quad (5)$$

α, β, γ and δ sum to 1 and are parameters used to weight the contribution given by each component of (5).

Finally, the relevance of a date t_j for a given document d is defined as the normalized sum of its relevance for each entity in the document. Thus we use the following ranking function for the candidate dates:

$$DRel(t_j, d) = \frac{\sum_{i=1}^{n} Conf(e_i) Rel(t_j, e_i)}{n} \quad (6)$$

where $Conf(e_i)$ is the level of confidence assigned by the NERD tool to e_i and n is the dimension (number of elements) of the set NE_d. We then take as estimated focus time of the document d the date in T_{cand} with the highest value of $DRel(t_j, d)$.

4 Experimental settings

4.1 Datasets

To evaluate the proposed approach we used the methodology described in [11] as a first baseline (referred to as B1 for now on). B1 was tested on document sets extracted from web sites and digital editions of history books, focusing on events related to five countries. Unfortunately original datasets are not available so we created our two test datasets accessing the same sources and following the steps described in [11]. We call the original datasets D1 and D2.

Web dataset. We collected paragraphs from three web sites reporting main events related to the five countries. The web sites we considered are the same as

in [11]: History Orb[9], History World[10], BBC Timelines[11], and Infoplease[12]. In order to reproduce a dataset as similar as possible to D1, we used the Wayback Machine[13] to access the snapshot of the websites recorded in January 2015, which is the date it was accessed to create D1. However, we found the average length of paragraphs (3.6 sentences) to be considerably smaller than that reported in D1 (18.3 sentences). As the performances of our method decrease with the inverse of the document length, due to the smaller number of entities matched in a short document, we believe this difference does not favour our approach against B1. The Web Dataset contains 1007 documents referring to events in the time span 1900-2015.

Books dataset. We collected text paragraphs from the two digital history books used in D2: Timeline of World History (Kerr, 2011) and Timelines of History (Ratnikas, 2012), following the procedure described in [11]. The resulting dataset has an average document length of 40 sentences (against 43 in D2) and the average event year is 1959 (against 1982 in D2). Such a difference should not favor our methods as - see [11] - the baseline method performs better for more recent events. The Books Dataset contains 747 documents referring to events in the time span 1900-2015.

4.2 Named entities spotting

Dandelion API, which we used to match named entities, provides a confidence score (from 0 to 1) indicating how sure the system is about the association. Errors in matching and disambiguating entities clearly affect final results. To mitigate this problem we filtered out entities with a confidence score below a threshold value th. We experimented with different values of th and manually checked the goodness of entities match on 20 randomly selected documents. We found 0.7 to be the optimal value. For this level of th, we found an average number of entities match per document of 8.9 and a total number of entities match of 8958 in the Web dataset. The books dataset, where documents are in average longer, has 92.8 average entity matches per document and 69303 entity matches in total.

4.3 Evaluation measures and parameters settings

In order to compare the proposed approach with the baseline methodology we use average error to measure results. The estimation error is computed as the absolute value of the difference between the estimated focus time and the ground truth. The ground truth date for a document can be, in our datasets, a single

[9] http://www.historyorb.com
[10] http://www.historyworld.net
[11] http://www.bbc.co.uk/history
[12] http://www.infoplease.com
[13] http://archive.org/web/

year or a range of years. In the latter case, as done in [11], the estimation error is calculated as the distance between the estimated year and the closest boundary of the range, or zero if the estimated year is included in the range. We thus adopt the following formula to represent the error $e(t)$:

$$e(t) = \begin{cases} min\{|t_b - t|, |t - t_e|\} & \text{if } t \notin [t_b, t_e], \\ 0 & \text{otherwise} \end{cases}$$

We then compute the average error over all the documents in the datasets.

In order to tune the parameters of our algorithm we randomly selected 10% of the document from both our datasets and used the remaining documents as a test set. We experimented on the training set with several different parameter settings, then we evaluated our method on the test set using the one providing the lowest average error. The selected configuration is the following: $\alpha, \gamma = 0.1666; \beta, \delta = 0.3333$.

In addition to the average error, we also measured precision, that is the number of documents correctly annotated ($e(t) = 0$) with respect to the ground truth date, divided by the total number of documents.

5 Results and discussion

In Table 1 we show the results of our experiments, where our method is referred to as *BOE* (Bag of Entities). In the table we compare them with two baselines: B1 and B2. B2 uses explicit dates mentioned in the documents. Given that we are interested in processing each document independently[14], we simply calculated the explicit-dates score by assigning $e(t) = 0$ if the document contains the ground truth year (or a year in a ground truth time span). Otherwise we computed $e(t)$ considering the date in the document which is closest to the ground truth as the estimated focus time.

Data	Method	Avg. error(years)	Avg. error(years)+	Prec.(%)	Prec.+(%)	Failed(%)
Books	B1	16.1	16.1	-	-	0.0
	B2	11.6	22.4	39.8	28.6	20.1
	BOE	**8.8**	**9.1**	**66.4**	**66.0**	0.6
Web	B1	20.2	20.2	-	-	0.0
	B2	7.5	40.1	58.3	13.2	76.8
	BOE	15.7	**16.4**	32.8	**32.2**	1.8

Table 1. Average error and precision of our best parameters settings vs. baseline and random estimation.

As shown, our method outperforms B1 with respect to average error. We remind the reader, however, that we ran the algorithm on datasets that are

[14] i.e. we do not consider co-occurrences of dates as well as of entities in the test corpus for our estimations

not identical to the ones used to evaluate B1. Even if they have been collected from the same sources and following the same methodology, our datasets contain shorter documents. As expected our method performs better on longer text (e.g. the books dataset), where more named entities can be spotted. No evaluation of B1 is provided in [11] with respect to precision. With our method the precision for book dataset is relatively high, while it is substantially lower in the web dataset, due to the shortest length of the documents, which are often composed by a short single sentence. This is also the reason why the number of non estimated documents (failed column in Table 1) is considerably higher in this dataset. Non estimated documents are usually very short documents where the entities spotted have a low level of confidence. An example is *Japanese government of Imukai forms*, where the two entities *Japanese government* and *Imukai* are matched with a confidence of respectively 0.44 and 0.26. On the other hand, B1 estimates all documents whereas B2, even though it provides a low average error (especially in the web dataset) fails to estimate more than 76% of the documents as few explicit dates are present on average in our test documents. To take into account failed estimations, in the columns marked with + we report the average error and the precision on the whole collection, considering non estimated ones. For such documents we have set the estimation error to half of the corpus time interval, which is the error we would obtain by arbitrarily choosing the middle of the interval as focus time.

Data	Method	Avg. error	Avg. error GT_{date}	Avg. error GT_{int}	Precision	Precision GT_{date}	Precision GT_{int}
Books	BOE	**8.8**	**6.2**	**18.6**	**66.4**	**75.9**	31.5
	wiki-only	13.5	11.8	19.8	54.0	58.0	**39.2**
	abst-only	14.7	11.9	25.0	54.9	62.4	27.3
	triples-only	24.9	25.9	21.4	6.1	4.6	11.9
	periods-only	11.5	8.6	22.0	46.8	55.6	14.7
Web	BOE	**15.6**	**16.5**	**1.3**	**32.8**	**29.7**	**84.0**
	wiki-only	20.5	21.3	6.7	31.4	28.9	74.0
	abst-only	21.7	22.4	9.4	26.0	23.8	64.0
	triples-only	22.8	23.3	14.8	14.4	12.6	44.0
	periods-only	22.2	23.0	9.9	7.7	6.3	32.0

Table 2. Compared evaluation of the different components of the proposed method.

In Table 2 we compare the results obtained with three different settings, each considering only the contribution of one element of the formula 5. The method marked as *wiki-only* weights time associations based only on the appearance of a date in the Wikipedia article ($\alpha = 1; \beta, \delta, \gamma = 0$). Similarly, *abst-only* considers only the abstract from DBpedia ($\beta = 1; \alpha, \delta, \gamma = 0$), *triples-only* considers only explicit dates found in DBpedia triples ($\gamma = 1; \alpha, \beta, \delta = 0$), and, finally, *periods-*

only scores dates only with respect to their inclusion in entity related time intervals detected in DBpedia ($\delta = 1; \alpha, \beta, \gamma = 0$).

We measure average error and precision with respect to the two distinct types of ground truth we have: single years and time spans. Precision in single years in books dataset is 75.6% while for time span is only 31.5%. This is unexpected behaviour, as in general we would expect it to be easier to match a period than a precise year. However, we notice that in the book dataset documents marked with a ground truth time span are rare (21 out of 747) and the intervals are quite short (3 years on average). On the other hand, in the web dataset, where we have 166 documents out of 1007 with a ground truth time span and the average length of such intervals is around 10 years, the precision is considerably higher for time spans than for single dates. Results clearly indicate that the combination of the different contributions increases the performance both in average error and precision when compared to single components, with the only exception of *wiki-only* which provides better precision in the books dataset in the case of time spans (GT_{int}).

6 Conclusions and outlook

In this paper we proposed a methodology for estimating focus time of a document based on named entities and temporal information extracted from Wikipedia and DBpedia. The methodology is designed to provide unsupervised and domain agnostic focus time estimation.

Results are encouraging and demonstrate that leveraging bag-of-entities is a good strategy for addressing focus time estimation.

In the future, we are planning to investigate how the approach could be integrated with [11] and with those based on explicit temporal expressions to increase performances, as well as evaluating the use of additional temporal data sources, e.g. YAGO.

Finally we remark that the methodology evaluated here performs single year focus time estimation only. Estimating the focus time as intervals of dates, as done in [11], as well as investigating higher granularity than years, is left for future works.

7 Acknowledgments

A special thanks goes to SpazioDati[15] for supporting our research by granting access to the Dandelion API. This work was supported by the GramsciSource FIRB project[16] funded by the Italian Ministry of Education, Universities and Research (MIUR).

[15] http://www.spaziodati.eu/
[16] http://gramsciproject.org

References

1. Alonso, O., Gertz, M., Baeza-Yates, R.: Clustering and exploring search results using timeline constructions. In: CIKM'09 (2009)
2. Campos, R., Dias, G., Jorge, A., Jatowt, A.: Survey of temporal information retrieval and related applications. ACM Computing Surveys 47(2) (2014)
3. Campos, R., Jorge, A., Dias, G., Nunes, C.: Disambiguating implicit temporal queries by clustering top relevant dates in web snippets. In: IEEE/WIC/ACM International Conference on Web Intelligence, WI'12 (2012)
4. Caputo, A., Basile, P., Semeraro, G.: Boosting a semantic search engine by named entities. In: ISMIS'09 (2009)
5. Cornolti, M., Ferragina, P., Ciaramita, M.: A framework for benchmarking entity-annotation systems. In: WWW'13 (2013)
6. Ferragina, P., Scaiella, U.: Tagme: On-the-fly annotation of short text fragments (by wikipedia entities). In: CIKM'10 (2010)
7. Garcia, M., Rodriguez, R., Anido Rifon, L.: Biomedical literature classification using encyclopedic knowledge: A wikipedia-based bag-of-concepts approach. PeerJ 2015(9) (2015)
8. Hoffart, J., Suchanek, F., Berberich, K., Weikum, G.: Yago2: A spatially and temporally enhanced knowledge base from wikipedia. Artificial Intelligence 194 (2013)
9. Hu, J., Fang, L., Cao, Y., Zeng, H.J., Li, H., Yang, Q., Chen, Z.: Enhancing text clustering by leveraging wikipedia semantics. In: SIGIR'08 (2008)
10. Jatowt, A., Au Yeung, C.M., Tanaka, K.: Estimating document focus time. In: CIKM'13 (2013)
11. Jatowt, A., Au Yeung, C.M., Tanaka, K.: Generic method for detecting focus time of documents. Information Processing and Management 51(6) (2015)
12. Kanhabua, N., Nrvg, K.: Improving temporal language models for determining time of non-timestamped documents. In: ECDL'08 (2008)
13. Kuchar J., K.T.: Bag-of-entities text representation for client-side (video) recommender systems. Workshop on Recommender Systems for Television and online Video (RecSysTV), at RecSys'14 (2014)
14. Martin, P., Spaniol, M., Doucet, A.: Temporal reconciliation for dating photographs using entity information. In: Workshop on Exploiting Semantic Annotations in Information Retrieval, at CIKM'15 (2015)
15. Mihalcea, R., Tarau, P.: Textrank: Bringing order into text. In: EMNLP'04 (2004)
16. Rizzo, G., Erp, M., Troncy, R.: Benchmarking the extraction and disambiguation of named entities on the semantic web. In: ESWC'14. Reykjavik, Iceland (2014)
17. Scaiella, U., Prestia, G., Del Tessandoro, E., Ver, M., Barbera, M., Parmesan, S.: Datatxt at #microposts2014 challenge. In: Workshop on Making Sense of Microposts, at WWW'14 (2014)
18. Stilo, G., Morbidoni, C., Cucchiarelli, A., Velardi, P.: Capturing users' information and communication needs for the press officers. In: Workshop on Social Media for Personalization and Search, at ECIR'17 (2017)
19. Strtgen, J., Alonso, O., Gertz, M.: Identification of top relevant temporal expressions in documents. In: Temporal Web Analytics Workshop, at TempWeb'12 (2012)
20. Strtgen, J., Gertz, M.: Multilingual and cross-domain temporal tagging. Language Resources and Evaluation 47(2) (2013)
21. Wang, F., Wang, Z., Li, Z., Wen, J.R.: Concept-based short text classification and ranking. In: CIKM'14 (2014)