IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received October 11, 2018, accepted November 13, 2018, date of publication November 23, 2018,
date of current version December 27, 2018.

Digital Object Identifier 10.1109/ACCESS.2018.2882988

Section-Based Focus Time Estimation of News Articles

SHAFIQ UR REHMAN KHAN, MUHAMMAD ARSHAD ISLAM, (Member, IEEE), MUHAMMAD ALEEM, MUHAMMAD AZHAR IQBAL, AND USMAN AHMED
Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan
Corresponding author: Shafiq Ur Rehman Khan (shaffiqmasud@gmail.com)

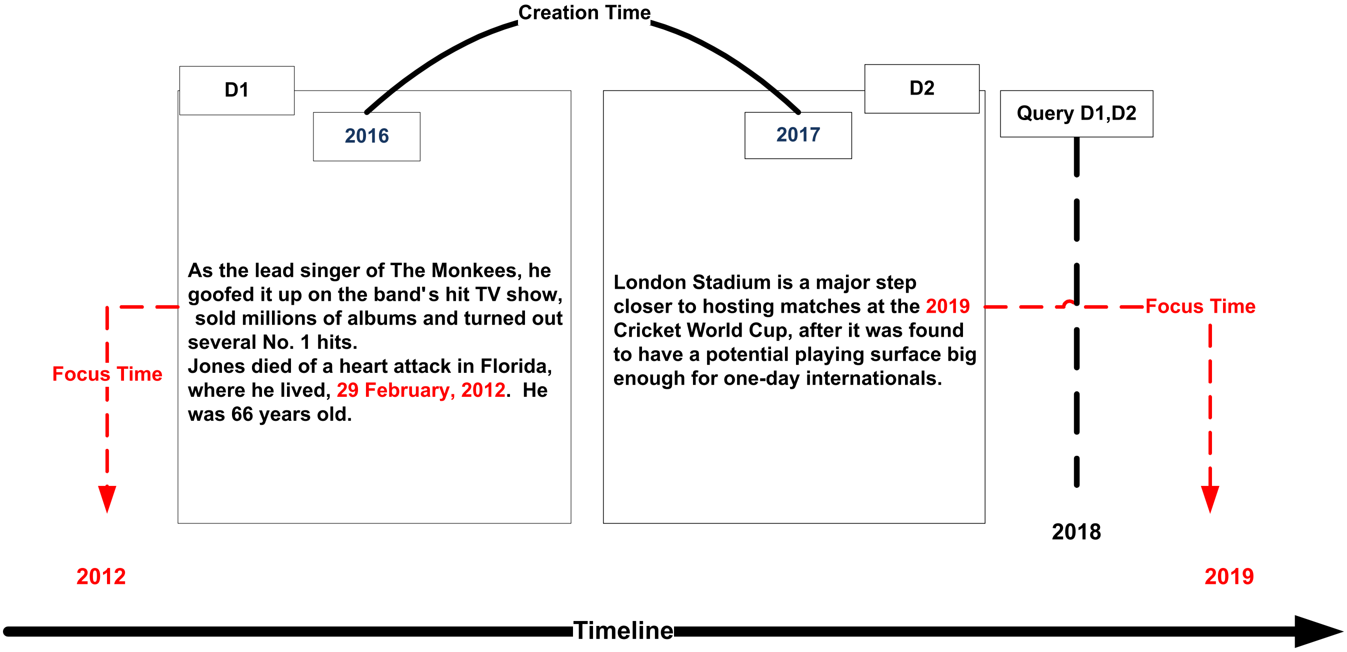
ABSTRACT Information retrieval systems embed temporal information for retrieving the news documents related to temporal queries. One of the important aspects of a news document is the *focus time*, a time to which the content of document refers. The contemporary state-of-the-art does not exploit focus time to retrieve relevant news document. This paper investigates the inverted pyramid news paradigm to determine the focus time of news documents by extracting temporal expressions, normalizing their value and assigning them a score on the basis of their position in the text. In this method, the news documents are first divided into three sections following the inverted pyramid news paradigm. This paper presents a comprehensive analysis of four methods for splitting news document into sections: the paragraph-based method, the words-based method, the sentence-based method, and the semantic-based method (SeBM). Temporal expressions in each section are assigned weights using a linear regression model. Finally, a scoring function is used to calculate a temporal score for each time expression appearing in the document. These temporal expressions are then ranked on the basis of their temporal score, where the most suitable expression appears on top. The effectiveness of the proposed method is evaluated on a diverse dataset of news related to popular events; the results revealed that the proposed splitting methods achieved an average error of less than 5.6 years, whereas the SeBM achieved a high precision score of 0.35 and 0.77 at positions 1 and 2, respectively.

INDEX TERMS Information retrieval, temporal information retrieval, focus time, inverted pyramid, news retrieval.

I. INTRODUCTION

While reading the news pertaining to the court judgment for compensation in 2017 for the Deepwater Horizon Spill (BP oil spill) -in the Gulf of Mexico, various questions crop up as a natural corollary in the minds of newsreaders such as, *When did the oil spill start? What were the reasons behind such an industrial disaster? Who was the president of BP Oil in 2011?* All of these questions focus on a particular time span when the incident occurred. Such types of information requirements are referred as temporal information needs. For instance, in the context of aforementioned questions, the newsreaders are interested in the news documents that contain information about the events (BP oil spill) occurred in 2011. To address such sort of queries, Information Retrieval (IR) systems that consider the news focused time for user temporal queries could assist in fulfilling the readers information needs.

In news documents, time is represented in the form of temporal expression, like calendar dates, or duration of time intervals [1]–[3]. Temporal expressions are classified into two broad types: explicit and implicit [4]. The former refers to a specific point in time which can be mapped directly to a date or a year [5], for instance August 14, 2014. Implicit temporal expressions describe some event without explicitly mentioning the time instant, for example Labor day, Christmas etc. Studies have shown that approximately 13.8% of the user queries contain explicit time expressions while 17.1% contain implicit time expressions [6]–[8], which are approximately trillion of temporal queries annually. Consequently, the Temporal Information Retrieval (TIR) has received significant attention from the research community in the recent years [9]. Plethora of studies have been conducted with an intention of satisfying the temporal information needs of users specified through temporal queries [10], [11]. Rapid increase in data volume (big data) [12] and web users make information retrieval a challenging task. Traditionally, IR systems (like search engines) mostly emphasize textual relevance whereas TIR systems consider both the textual



As the lead singer of The Monkees, he goofed it up on the band's hit TV show, sold millions of albums and turned out several No. 1 hits.
Jones died of a heart attack in Florida, where he lived, 29 February, 2012. He was 66 years old.

London Stadium is a major step closer to hosting matches at the 2019 Cricket World Cup, after it was found to have a potential playing surface big enough for one-day internationals.

FIGURE 1. Time differences between query time, creation time, and focus time of a news document

relevance and the temporal relevance of the query to retrieve the most temporally and textually relevant documents. The time dimension is considered in a numerous information retrieval processes including document pre-processing, ranking/retrieval models, and query processing.

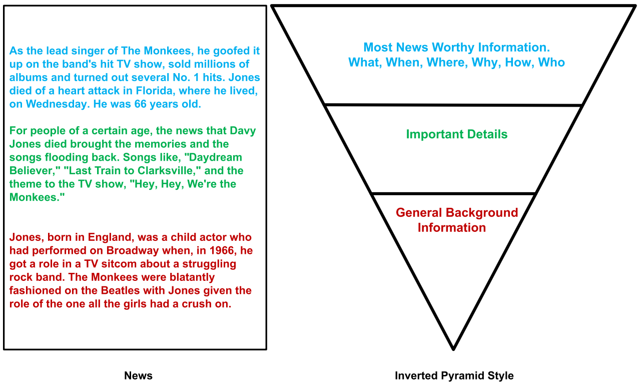
To address the temporal queries, IR systems should retrieve the documents that match the intended time scope of the temporal queries. One simple approach to retrieve temporally relevant documents is to consider the documents creation time. However, the suitability of using the creation time is questionable since: 1) the document creation time may be different to the published time; and, more importantly, 2) the focus time of the document may not match the creation time. The document focus time is defined as the time referred by the content of the document [13]; this is particularly important when the user is interested in a temporal focus of the document with the interest in some past or future event. Such a scenario is presented in Figure 1, where two news documents D1 and D2 are created in the years 2016 and 2017, respectively. Contemplate a scenario, where a user poses queries in 2018 with an intention to search news related to ‘Davy Jones’ death in 2012 (D1) and ‘Cricket World Cup 2019’ (D2). In such scenarios, the focus time is more important than the creation or publication time of the documents.

One of the important functionality of a search engine is news retrieval. News search systems constantly index the news from different sources worldwide and facilitate the users searching for news. Creation time plays an essential role in retrieving a news document; however, we argue that most of the time user is interested in the focus of news rather than its creation time. As best of our knowledge, focus time has not yet been considered as per its importance for treating the temporal queries in IR systems. This issue has grabbed scant attention in the scientific community.

The inverted pyramid news structure is the most common reporting style of English news [14]. Carole Rich [15] define the inverted pyramid style in the following terms:

’The most common type of lead on the hard-news story is called a “summary lead” because it summarizes the main points about what happened. It answers the question who, what, when, where, why and how. The rest of the story elaborates on what, why and how.

According to above syntax, the most important, newsworthy, and relevant information is at the top, followed by the less relevant, with the least important at the bottom [16] as illustrated in Figure 2. This structure motivates us to divide a


As the lead singer of The Monkees, he goofed it up on the band's hit TV show, sold millions of albums and turned out several No. 1 hits. Jones died of a heart attack in Florida, where he lived, on Wednesday. He was 66 years old.

For people of a certain age, the news that Davy Jones died brought the memories and the songs flooding back. Songs like, "Daydream Believer," "Last Train to Clarksville," and the theme to the TV show, "Hey, Hey, We're the Monkees."

Jones, born in England, was a child actor who had performed on Broadway when, in 1966, he got a role in a TV sitcom about a struggling rock band. The Monkees were blatantly fashioned on the Beatles with Jones given the role of the one all the girls had a crush on.

Most News Worthy Information.
What, When, Where, Why, How, Who
Important Details
General Background Information
News
Inverted Pyramid Style

FIGURE 2. Inverted pyramid news paradigm..

news article into three sections for the purpose of identification of accurate focus time. Hence the main research question investigated in this study is:

RQ: *How accurately can we determine the focus time of news documents by assigning weights to temporal expressions located in implicit sections of documents?*

To the best of our knowledge, this study is the first attempt to investigate the inverted pyramid news paradigm for focus time detection. The main contribution of this work is a novel approach for section-based ranking of temporal expressions to determine the focus time of the document. In this approach, the news document is first divided into three sections, and then a temporal score is assigned to each temporal expression based on their position in the text. Four methods - the Paragraph Based Method (PBM), the Words Based Method (WBM), the Sentence Based Method (SBM), and the Semantic Based Method (SeBM) - are used to divide the news document into three logical sections. Temporal weights are assigned to each logical section, and temporal scores are calculated for each temporal expression using scoring function. These temporal expressions are then ranked in such a way that the top one is the most suitable candidate for focus time.

Note: From here on, "news article" and "document" are used interchangeably - they carry the same meaning. Similarly, "focus time" and "focus year" are the same as we set the time granularity to a year.

The rest of the paper is structured as follows. After reviewing related work in the next section, we present the news document pre-processing, storing of the temporal information, and the approach to divide news document into three logical sections in Section III. The gold standard (used for evaluation) construction and the scoring function for ranking temporal expressions are described in Section IV, followed by results and discussion in Section V. Finally, Section VI concludes this research work and present future directions.

II. RELATED WORK

Temporal information retrieval is an emerging sub-field of information retrieval [8], [17], [18], and there are many information retrieval applications on the Internet that use time as a primary feature for searching [19], [20]. It aims to satisfy users temporal information needs by considering temporal relevance along with the textual relevance. The creation time of a document is usually important to retrieve a temporal document, and in most commercial search engines, results are ranked based on the document creation time. However, there are two problems associated with the document creation time: 1) the creation date of documents is not always available; and 2) the document creation time may not represent the focus time of the document. Ranking documents by their creation time may decrease the effectiveness of IR system when the user is not interested in the creation time but the focus time of the document - for example, a document created in 2015 discussing the FIFA WORLD CUP 2022 event.

Several approaches have been proposed to estimate the creation time of non-time-stamped documents, and the process has been named "Document Dating". Alonso [21] classifies this work into content-based and non-content-based methods. In the content-based method, the content of the document is used for document dating; this needs a dependent time-stamped document collection in order to create a model. On the other hand, non-content-based document dating uses external information; the major shortcoming of such methods is the lack of availability and accuracy of external sources. Earlier work by De Jong *et al.* [22] used a statistical language model to estimate document creation time. In this approach, reference data is partitioned into several time granularities and temporal language models are constructed for each partition. The language model of the undated document is then compared with the temporal language model of each partition. Kanhabua and Nørvåg [23] extended this model using temporal entropy, Google Zeitgeist, and semantic pre-processing. In another work, Filannino and Nenadic [24] extracted the temporal expressions from document text and constructed a time line associated with a specific entity (person pages from Wikipedia), predicting its upper and lower boundaries. Niculae *et al.* [25] employed a statistical model to predict the document creation date using documents presented in three languages: English, Portuguese, and Romanian.

One of the most significant works in this field, by Jatowt *et al.* [13], estimated document focus time through word time association. Terms are extracted from news article of different years, and these words are then associated with a time. If the document has many words associated with a certain time period *t*, then the document has a strong association with time period *t*. Another work by Spitz *et al.* [26] presented a graph based ranking model wherein the set of words relevant to certain time periods is determined. Authors suggested that the more often a term appears with a temporal expression at the sentence level in the document, the more likely it is that the term and date are related. Our work is different in such a way that we consider implicit and explicit temporal expressions frequency in specific section of a news article following the inverted pyramid paradigm hypothesis where certain parts of news article carry different level of useful information. We assign different weights to three sections of a news article based on the assessed importance of the sections for determining the focus time.

III. METHODOLOGY

The overall scheme of focus time detection is presented in Figure 3. Document pre-processing, logical section creation and document temporal profiling processes are elaborated in this section. Whereas, the dataset acquiring process, data annotation and temporal ranking function are discussed in Section IV.

A. DOCUMENT PRE-PROCESSING

This step includes the standard document pre-processing involved in information retrieval processes, such as tokenization, stop words removal, stemming, and calculating

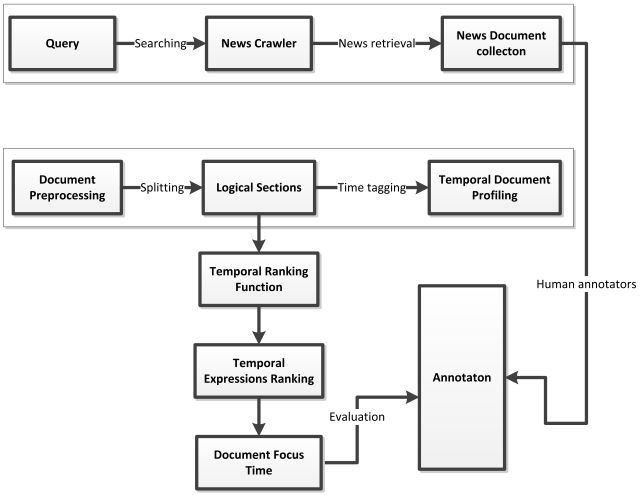


FIGURE 3. **The proposed system architecture used in this study.**

term frequencies. This standard pre-processing procedure is followed by the identification, extraction, and normalization of temporal expressions.

B. LOGICAL SECTIONS

We divide the news documents into three logical sections using four alternate methods (WBM, SBM, PBM and SeBM) to analyze the potential of each section in identifying the focus time. These four methods are delineated below.

1) WORD BASED METHOD

In this method, the news document is divided into three sections based on words/terms count, where each section contains a specific proportion of the total words. The first section contains 50% of all the words, followed by section 2 containing 30%, and the remaining 20% of the words are placed in the third section. The rationale for such a division is that the first section is the most important as it contains the most useful information about the news. To minimize the chance of losing the important information, we set the size of the first section reasonable large. The reason for the small size of the third section is that the last section of a news article typically contains little background information so we keep it small.

2) SENTENCE BASED METHOD

The news documents are split into single sentences before dividing the text into three sections. The first section contains 40% of all the sentences, whereas 40% and 20% of sentences are allocated to sections 2 and 3, respectively.

3) PARAGRAPH BASED METHOD

In this method, the first section contains the first paragraph of the news document, whereas the remaining paragraphs are assigned to sections 2 and 3 based on a 3:2 ratio respectively. The reason for assigning first paragraph to first section is that the first paragraph contains abstract information about the event.

4) SEMENTIC BASED METHOD

In this method, the news documents are divided into three sections based on criteria fulfilling the inverted pyramid concept. As shown in Figure 2, the first section answers the *what*, *when*, *where*, and *who*. In order to extract information for aforementioned questions, the content of news document is first searched for the phrases that represent *what*, *when*, *where*, and *who*. *What* refers to the question "what is the actual event?", *when* determines the time of the event, *where* represents the geographical location of the event, and *who* some person or organization involved in the event.

For the first aspect (*what*), the title of the article contains a description of the event so keywords are extracted from the title. To extract information about *when*, we apply a temporal tagger to the text, which identifies and normalizes temporal expressions. Finally, Stanford Name Entity Recognition (NER) [6] is used to tag geographical locations, persons, and organizations to answer *where* and *who*. This method works in such a way that our system searches for title keywords, time, geographical location, and entity; the first section boundary is drawn where these appear for the first time in the text. The remaining sections are created on the basis of a 3:2 ratio: the remaining 60% of the text in section 2 and 40% in section 3.

C. DOCUMENT PROFILING

For temporal expression identification, extraction, and normalization, we use HeidelTime [27]. HeidelTime is a rule-based temporal expression extraction and normalization tool that mainly uses a regular expression for temporal expression extraction and knowledge resources as well as linguistic clues for their normalization. HeidelTime uses creation time as a reference when normalizing the temporal expression. For each splitting method $sm = \{WBM, SBM, PBM, SeMB\}$, such temporal information is stored in a database where each record presents information about a single temporal expression:

$$te_n = \{doc : id, sid, ae, ne, nd, nm, ny\} \tag{1}$$

te_n represents n^{th} temporal expression, $doc : id$ is the document identification, sid is section id where the te_n appears, ae is the actual expression, ne is normalized expression; nd, nm and ny show the normalized day, month, and year respectively. This information is then used to construct the documents temporal profiles, represented as:

$$tp_d = \{doc : id, tid, ct, ny_{s1}, ny_{s2}, ny_{s3}\} \tag{2}$$

Where, tp_d is temporal profile of document d containing information about document $doc : id$, temporal expression identification tid, creation time of document ct and ny_{s1}, ny_{s2}, and ny_{s3} are the normalized years in section 1, section 2, and section 3, respectively.

IV. EXPERIMENTAL SETUP

The motivation behind the experiments conducted in this paper is to assess the focus time of the news documents.

TABLE 1. **Temporal queries used to crawl the news documents from Google news archive.**

Q.No	Query	Q.No	Query	Q.No	Query	Q.No	Query	Q.No	Query
Q1	Ambassador Steven Death, 2012	Q8	David Cameron Resignation, 2016	Q15	Hurricane Sandy, 2012	Q22	Moscow Terror Attack, 2010	Q29	Tunisia Revolutions, 2011
Q2	Athens Wildfire, 2009	Q9	Kashmir Earthquake, 2005	Q16	South Sudan Independence, 2011	Q23	Cyclone Nargis, 2008	Q30	Robbin Williams Death, 2014
Q3	Baltimore Riots, 2015	Q10	Fidel Castro Retirement, 2008	Q17	London Bombing, 2005	Q24	Pakistan Flood, 2010	Q31	Saddam Hussein Execution, 2006
Q4	Benazir Assassination, 2007	Q11	FIFA Football World Cup, 2022	Q18	Madrid Terrorist Attacks, 2004	Q25	Pervaiz Musharraf Resignation, 2008	Q32	Sochi Olympics, 2014
Q5	Pope Benedict XVI, 2005	Q12	Fukushima Disaster, 2011	Q19	MH370 Disappearance, 2014	Q26	Prince Charles Wedding, 2005	Q33	Steve Jobs Death, 2011
Q6	BP Oil Spill, 2010	Q13	Haiti Earthquake, 2010	Q20	Michael Jackson Death, 2009	Q27	Prince William Wedding, 2011	Q34	Switzerland Joined UN, 2002
Q7	Cricket World cup, 2019	Q14	Hurricane Katrina, 2005	Q21	Mike Tyson "The Bite Fight",1997	Q28	Rayan Dunn Death, 2011	Q35	Volkswagen Scandal, 2015

The reason for selecting the news documents is twofold: the news documents have creation time and secondly, the news documents are enriched with temporal expressions, which are very interesting for this study.

As mentioned earlier, the contemporary literature has merely concentrated on focus time. Therefore, there is a lack of gold standard dataset that can be employed to evaluate the outcomes of proposed scheme. Therefore, we conducted a user study to evaluate the effectiveness of the proposed methods. For this, news documents are distributed among the post graduate students. In the rest of this section, process of gold standard dataset construction is presented(in section A), followed by the scoring function for ranking (in Section B).

A. DATASET

There is no standard focus time benchmark available to test our approaches. To construct the dataset, we extracted the news documents devoted well-known events form Google News, as shown in Table 1. Google News tool is a search engine that particularly extracts the news documents. It provides a platform to the user to search required news documents using some searching criteria. Google News is a custom Internet newspaper that contains articles from 4,500 different news sources and adopts all search functions of Google. We built a crawler to extract the news form Google news. It performs searching according to query of selected events and collects the documents corresponding to each event.

In order to retrieve most relevant documents, we use explicit temporal queries $Q_t = \{q_{text}, q_{time}\}$ comprises of two parts: textual part q_{text} and temporal part q_{time}, where $q_{text} = \{w_1, \ldots, w_m\}$ and $q_{time} = \{t_{year}\}$. The textual part q_{text} comprises of query terms (i.e.,event name) and the temporal part t_{year} is the year when the event occurred. Such queries are normally referred as explicit temporal queries. Queries that explicitly mention time, capture the real world meaning of time [28]. For instance, to collect relevant news documents pertaining to an event of Prince Charles wedding, the query is ''Prince Charles Wedding 2005''.

The top_k news articles ($k = 100$ ranked by the Google news search are crawled for each event. We collected a total of 3500 news documents against 35 queries. A gold standard is built by relying on human judgments in identifying the actual focus time from news documents. Total of 3500 news documents were assigned 70 post-graduate students. Each participant was assigned with 100 news documents about a specific event (query), and were asked to label each document as relevant or irrelevant according to the given query. Thus, for each event the 100 news documents are labeled by 2 participants. Relevance of a document to the query obviously ensures that the document relates to a corresponding event (i.e., event presented in the query). If annotators found that a document contains the information about event presented in the query, then they marked them as relevant, otherwise non-relevant.

The participants were requested to provide the reason for their judgment i.e., why they thought the document to be relevant or irrelevant. Such method ensures that the annotator read and understand the document properly. Finally, we consider those documents to be relevant where both the participants are agreed upon. Total of 918 out of 3500 news documents, were marked as relevant by the human annotators. The relevant documents against each individual query in the dataset is presented in Figure 4. Figure 5 presents the statistics of relevant and non-relevant documents in the dataset.

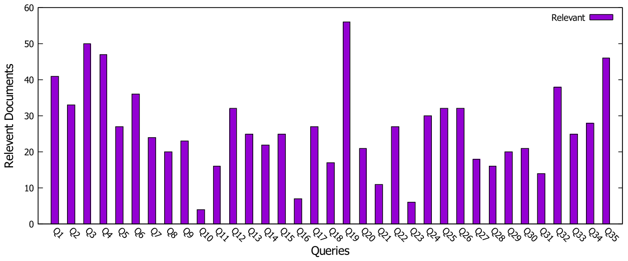


FIGURE 4. **Relevant document distribution over individual queries in the dataset..**

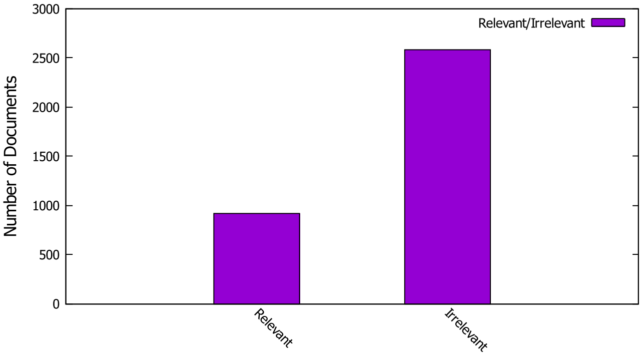


FIGURE 5. **Relevant and non-relevant news documents in the corpus.**

B. TEMPORAL SCORING FUNCTION

The Temporal Scoring Function (TSF) assigns a score to the temporal expressions (years) by analyzing the position of expression in the text of news document. The temporal scoring function is defined as.

$$ts(te) = \alpha_1(\sum te_{s1}) + \alpha_2(\sum te_{s2}) + \alpha_3(\sum te_{s3}) \quad (3)$$

Where $ts(te)$ is the temporal score of expression te, $\sum te_{s_i}$ present the count of temporal expression te in each of the three sections s1, s2 and s3. α_1, α_2 and α_3 are temporal weights (constants) that are assigned to each temporal expression appears in each of the three logical sections. The weights are calculated using multi-linear regression model by using the temporal characteristics of 918 relevant documents. The temporal expression occurred in section 1 attained temporal weight of 0.9, the highest weight; the temporal weight then decreased to 0.6 and 0.3 in the subsequent section 2 and 3, respectively. These weights represent the section importance in terms of their informativeness, and hence receive more weight than those sections containing less information. After scoring each temporal expression in the document, these scores are ranked in descending order according to their temporal score.

C. EVALUATION

To evaluate the proposed methods for document splitting and scoring function, the following two evaluation measures are used.

1) PRECISION

The performance of the splitting methods and scoring function is evaluated using precision- a standard evaluation measure used in IR studies. We considered precision at position 1 (P@1) and precision at position 2 (P@2). Such measures present the number of documents for which the focus time is correctly determined at rank position 1 and 2. The precision is defined as:

$$P@n = \frac{CDF_n}{N_D} \quad (4)$$

Where n presents the rank $n \in \{1, 2\}$, CDF_n is the count of document for which the actual focus time is ranked at position n and N_D represents the number of documents in the dataset.

2) AVERAGE ERROR YEAR

The second evaluation measure is *Average Error Years (AEY)*. AEY is the mean difference between the actual focus time and the estimated focus time [13]. An error year can be calculated using the following expression:

$$e(y) = \begin{cases} |t_{fy} - t_{py}|, & \text{IF } t_{py} \notin t_{fy} \\ 0, & \text{otherwise} \end{cases} \quad (5)$$

Here $e(y)$ is the error year estimation, t_{fy} is the focus time (year) in the ground truth, t_{py} is the time (year) calculated by scoring function. The value of error years $e(y)$ is the difference between predicted focus time t_{py} and the actual focus time t_{fy}.

V. RESULTS AND DISCUSSION

The temporal score for each temporal expression (year) is calculated using Equation 3, and these expressions are ranked in descending order according to their corresponding scores. The higher the temporal score, the higher the rank of the temporal expression in the ranked list. The top ranked expression is assumed to be the best candidate for document focus time. After ranking the temporal expressions in descending order, we select the top two temporal expressions as the candidates for the focus time of the document. Document splitting methods have an impact on accurately estimating the focus time of a news document. The splitting methods and scoring function are evaluated using P@1 (Figure 6), P@2 (Figure 7) and average error years (Figure 9).

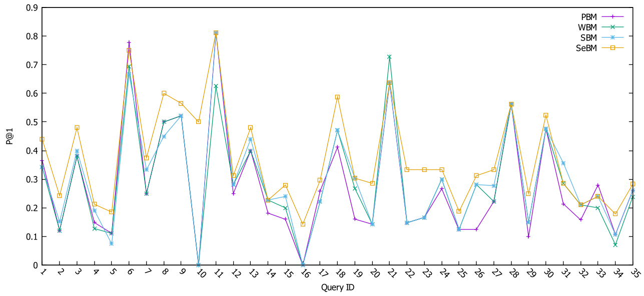


FIGURE 6. Precision achieved for temporal queries at position 1..

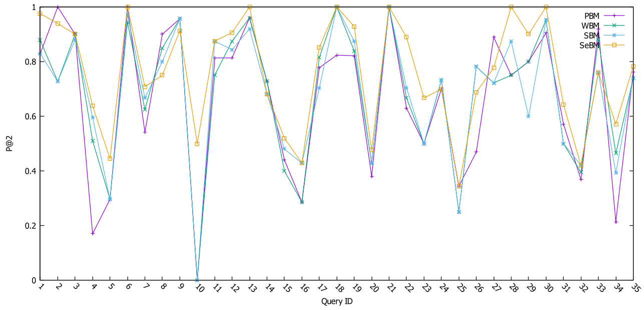


FIGURE 7. Precision achieved for temporal queries at position 2..

In Figure 6, the x-axis represents the query, and the y-axis presents the number of documents for which our proposed approach estimates the correct focus time at position 1. The colored lines present the splitting methods i.e.,orange = SeMB, blue = SBM, green = WBM, purple = PBM. Figure 6 illustrate P@1 score for individual query documents. The plot shows that SeBM achieves a high P@1 score (orange line) as compared to other splitting methods. In Figure 7, P@2 score is presented for the individual query documents, once again, the SeMB achieved high P@2 score as compared to other splitting methods.

Precision scores of 0.2756, 0.2846, and 0.3009 are achieved by PBM, WBM, and SBM, respectively at position 1 (P@1). Whereas, SeBM performed comparatively better than other three splitting methods by obtaining P@1 score of 0.3576, as illustrated in Table 2. The P@2 values achieved by PBM, WBM, SBM, and SeBM are 0.6815, 0.7099, 0.7077, and 0.7709, respectively. The performance of SeBM

TABLE 2. The precision achieved by scoring function at position 1 and 2 using the four splitting methods.

Method	P@1	P@2
PBM	0.2759	0.6815
WBM	0.2846	0.7099
SBM	0.3009	0.7077
SeBM	**0.3576**	**0.7709**

positively steadied at both P@1 and P@2 values. The results are illustrated in Figure 8.

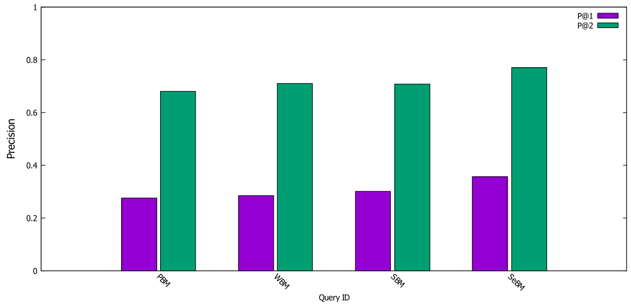


FIGURE 8. Precision at Position 1 and 2 achieved by the splitting methods.

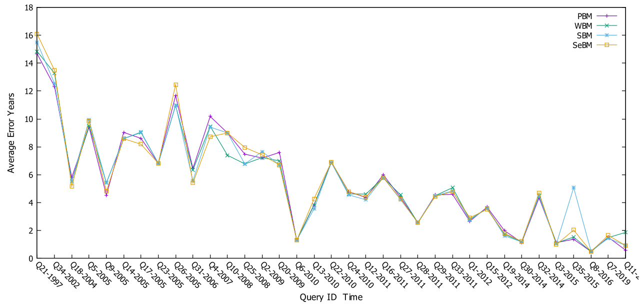


FIGURE 9. Year error distribution for individual temporal queries..

Turning now to the experimental evidence on error years estimation, Figure 9 presents the error years distribution for all the queries. The queries are arranged in chronological order (i.e. starting from the earliest year). The error years for queries "Mike Tyson 'The Bite Fight',1997", "Prince Charles Wedding, 2005", "Switzerland Joins UN, 2002", and "Pope Benedict XVI, 2005" are much higher than other queries. This is due to the difference in time between the event date and the query date (i.e., 2018), which are 21, 13, 16, and 13 years respectively. Less error years are observed for documents related to events that occurred near the query date. For example, "David Cameron Resignation, 2016", " FIFA Football World Cup, 2022" and "Robin Williams Death, 2014" are the events that occurred within less time interval, where the time difference (query time and event time) are 2, 4 and 4 years, respectively. The impact of event and query time difference on error years is presented in Figure 10, where less error years are observed for those events that occurred in a closer time span of the query time (2018). Various other pertinent reasons might be the popularity and the time span of event. For instance, the news about the disappearance of Malaysian airline flight "MH370" in 2014, which is still a mystery and its new updates still diffuse across several news platforms.

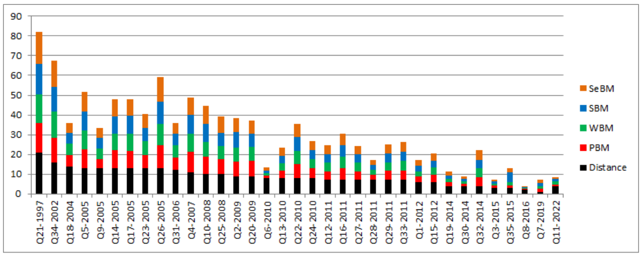


FIGURE 10. The impact of event and query time distance on error years.

In Figure 11, the average error years for all four proposed methods are illustrated. It is observed that the paragraph splitting method has the lowest average error years i.e., 5.51 followed by the words based splitting method with 5.541 error years; whereas, the semantic and sentence based splitting methods have attained average error of 5.54 and 5.56 years, respectively.

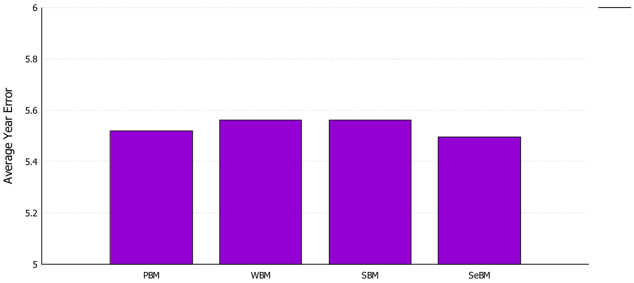


FIGURE 11. Average error years estimation for all the four splitting methods.

To recapitulate, temporal expressions in news documents are ranked in descending order based on a temporal score obtained by scoring function. The higher the temporal score, the higher the position of the temporal expression in the ranked list. The top two temporal expressions are considered for document focus time. Document splitting methods have an impact on accurately estimating the focus time of a news document. The results revealed that when the documents are split using SeBM, the scoring function accurately assesses focus time of 328 out of 917 documents at position 1. For SBM, the accurate focus time of 276 documents appeared at the top of list, whereas with WBM and PBM, the scoring function accurately estimated the focus time of 261 and 253 documents, respectively. For most of the documents, our proposed scoring function ranks the actual focus time at position 2 in the ranked list. The accurate assessment of focus time using SeBM, SBM, WBM, and PBM at position 2 are 707, 649, 651, and 625 documents respectively. Furthermore, the query and event time difference is also investigated with respect to values of precisions (i.e., P@1 and P@2). However, no significant impact of time difference (query time and event time) is observed in the values of precisions. The second evaluation measure used in this study is average error years, which is the difference between the estimated focus time and the actual focus time used by Jatowt *et al.* [13]. Using the temporal scoring function, the SeBM method has fewer average error years whereas the SBM and WBM have higher error years (see Figure 11).

VI. CONCLUSION AND FUTURE WORK

This study scrutinizes the potential of focus time for relevant news retrieval, which has been ignored by the existing state-of-the-art. This paper seeks to contribute new insight to the process of focus time assessment of news document using the inverted pyramid paradigm. For this purpose, we split the news articles into three sections using four methods (PBM, WBM, SBM, and SeMB). These news documents are then preprocessed and temporally annotated. The temporal profiles of the news documents are constructed and the temporal information is stored in a database. The temporal scoring function is used to calculate the score of each temporal expression in the news document and to rank these in such a way that the high scoring temporal expressions remain on the top of the list. In order to construct a gold standard, a user study is conducted by involving University students. The performance of the proposed scheme is evaluated using two evaluation methods. The first method uses precision at positions 1 and 2, whereas, the second method calculates the average error years between the actual focus time and estimated focus time. The evaluation results depicted that SeBM outperformed other splitting methods in terms of focus time detection. Using the scoring function and SeBM, a precision score of 0.35 is achieved, which means that for 35% of documents, the focus time is accurately estimated at position 1, whereas at position 2, 77% of documents are correctly labeled with focus time.

This research has opened various other directions that should be investigated in future. First of all, a better understanding of web news documents needs to be developed. For instance, a careful understanding of other news writing styles along with the inverted pyramid news paradigm. Moreover, the role of spatial references in news text might also play a role in estimating the focus time of the document. We believe that time and geographical location have a strong association when it comes to assessment of news document focus time.

REFERENCES

[1] R. Saurí, J. Littman, B. Knippen, R. Gaizauskas, A. Setzer, and J. Pustejovsky, "TimeML annotation guidelines," *TERQAS Annotation Working Group*, vol. 1, 2006.

[2] H. Holzmann, W. Nejdl, and A. Anand, "Exploring Web archives through temporal anchor texts," in *Proc. ACM Web Sci. Conf.*, Jun. 2017, pp. 289–298.

[3] E. Segev and A. J. Sharon, "Temporal patterns of scientific information-seeking on Google and Wikipedia," *Public Understand. Sci.*, vol. 26, no. 8, pp. 969–985, Nov. 2017.

[4] O. Alonso, R. Baeza-Yates, J. Strötgen, and M. Gertz, "Temporal information retrieval: Challenges and opportunities," in *Proc. 1st Temporal Web Anal. Workshop WWW*, 2011, pp. 1–8

[5] N. Kanhabua and K. Nørvåg, "Learning to rank search results for time-sensitive queries," in *Proc. 21st ACM Int. Conf. Inf. Knowl. Manage.*, Oct. 2012, pp. 2463–2466.

[6] D. Metzler, R. Jones, F. Peng, and R. Zhang, "Improving search relevance for implicitly temporal queries," in *Proc. 32nd Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2009, pp. 700–701.

[7] M. Shokouhi, "Detecting seasonal queries by time-series analysis," in *Proc. 34th Int. ACM SIGIR Conf. Res. Develop. Inf. Retr.*, Jul. 2011, pp. 1171–1172.

[8] S. Nunes, C. Ribeiro, and G. David, "Use of temporal expressions in Web search," in *Proc. Eur. Conf. Inf. Retr.* Springer, 2008, pp. 580–584.

[9] H. Wang, Q. Zhang, and J. Yuan, "Semantically enhanced medical information retrieval system: A tensor factorization based approach," *IEEE Access*, vol. 5, pp. 7584–7593, 2017.

[10] S. U. R. Khan, M. A. Islam, M. Aleem, and M. A. Iqbal, "Temporal specificity based text classification for information retrieval," *TURKISH J. Elect. Eng. Comput. Sci.*, 2018.

[11] M. Zahedi, A. Aleahmad, M. Rahgozar, F. Oroumchian, and A. Bozorgi, "Time sensitive blog retrieval using temporal properties of queries," *J. Inf. Sci.*, vol. 43, no. 1, pp. 103–121, Feb. 2017.

[12] C. Stergiou and K. E. Psannis, "Algorithms for big data in advanced communication systems and cloud computing," in *Proc. IEEE 19th Conf. Bus. Inform. (CBI)*, Jul. 2017, pp. 196–201.

[13] A. Jatowt, C. M. A. Yeung, and K. Tanaka, "Generic method for detecting focus time of documents," *Process. Manage.*, vol. 51, no. 6, pp. 851–868, Nov. 2015.

[14] E. A. Thomson, P. R. White, and P. Kitley, "'Objectivity' and 'hard news' reporting across cultures," *Journalism Stud.*, vol. 9, no. 2, pp. 212–228, Mar. 2008.

[15] C. Rich, *Writing and Reporting News: A Coaching Method*. Boston, MC, USA: Cengage Learning, 2015.

[16] H. Zhang and H. Liu, "Visualizing structural 'inverted pyramids' in English news discourse across levels," *Text Talk*, vol. 36, no. 1, pp. 89–110, Jan. 2016.

[17] R. Campos, G. Dias, A. M. Jorge, and A. Jatowt, "Survey of temporal information retrieval and related applications," *ACM Comput. Surv.*, vol. 47, no. 2, p. 15, Jan. 2015.

[18] N. Kanhabua, R. Blanco, and K. Nørvåg, "Temporal information retrieval," *Found. Trends Inf. Retr.*, vol. 9, no. 2, pp. 91–208, Jul. 2015.

[19] J. Murphy, N. H. Hashim, and P. O. Connor, "Take me back: Validating the wayback machine," *J. Comput.-Mediated Commun.*, vol. 13, no. 1, pp. 60–75, Oct. 2007.

[20] J. Hoffart, F. M. Suchanek, K. Berberich, and G. Weikum, "YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia," *Artif. Intell.*, vol. 194, pp. 28–61, Jan. 2013.

[21] O. R. Alonso, *Temporal Information Retrievel*. Davis, CA, USA: Univ. California, 2008.

[22] F. De Jong, H. Rode, and D. Hiemstra, "Temporal language models for the disclosure of historical text," in *Proc. 16th Int. Conf. Assoc. Hist. Comput.*, Sep. 2005, pp. 161–168.

[23] N. Kanhabua and K. Nørvåg, "Using temporal language models for document dating," in *Proc. Joint Eur. Conf. Mach. Learn. Knowl. Discovery Databases*. Springer, Sep. 2009, pp. 738–741.

[24] M. Filannino and G. Nenadic, "Mining temporal footprints from wikipedia," in *Proc. 1st AHA!-Workshop Inf. Discovery Text*, 2014, pp. 7–13.

[25] V. Niculae, M. Zampieri, L. Dinu, and A. M. Ciobanu, "Temporal text ranking and automatic dating of texts," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, vol. 2, 2014, pp. 17–21.

[26] A. Spitz, J. Strötgen, T. Bögel, and M. Gertz, "Terms in time and times in context: A graph-based term-time ranking model," in *Proc. 24th Int. Conf. World Wide Web*, May 2015, pp. 1375–1380.

[27] J. Strötgen and M. Gertz, "A baseline temporal tagger for all languages," in *Proc. Conf. Empirical Methods Natural Lang. Process.*, 2015, pp. 541–547.

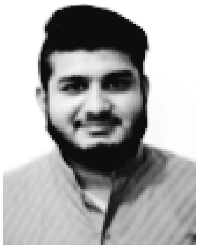
[28] N. Kanhabua, "Time-aware approaches to information retrieval," Ph.D. dissertation, Norwegian Univ. Sci. Technol., Feb. 2012.

SHAFIQ UR REHMAN KHAN received the B.S. degree in information technology from Gomal Univeristy, Dera Ismail Khan, Pakistan, in 2007, and the M.S. degree in software engineering from Bahria University, Islamabad, Pakistan. He is currently pursuing the Ph.D. degree in computer science with the Capital University of Science and Technology, Islamabad. His research interests are information retrieval, data mining, and machine learning.

MUHAMMAD ARSHAD ISLAM received the Ph.D. degree from the University of Konstanz, Germany, in 2011. His dissertation is related to routing issues in opportunistic network. He is currently an Assistant Professor with the Capital University of Science and Technology, Islamabad, Pakistan. His current research interests are related to MANETs, DTNs, social-aware routing, and graph algorithms.

MUHAMMAD ALEEM received the Ph.D. degree in computer science from Leopold-Franzens-University, Innsbruck, Austria, in 2012. He is currently an Assistant Professor with the Department of Computer Science, Capital University of Science and Technology, Islamabad, Pakistan. His research interests include parallel and distributed computing comprises programming environments, multi-/many-core computing, performance analysis, cloud computing, big data processing, and scheduling.

MUHAMMAD AZHAR IQBAL received the Ph.D. degree in communication and information systems from the Huazhong University of Science and Technology, Wuhan, China, in 2012. He is currently an Assistant Professor with the Computer Science Department, Capital University of Science and Technology, Islamabad, Pakistan. His research areas include coding-aware routing in vehicular ad hoc networks, energy-efficient MAC for wireless body area networks, large-scale simulation modeling, and analysis of computer networks in cloud.



USMAN AHMED received the B.Sc. degree (Hons.) in computer science from Heavy Industries Taxila Education City, Taxila, Pakistan. He is currently a Lecturer with the Department of Computer Science, University of Lahore Islamabad Campus, Islamabad, Pakistan. His current research interests include heterogeneous computing, natural language processing, and machine learning.