applied sciences

MDPI

Article

A Focused Event Crawler with Temporal Intent

Hao Wu [1] and Dongyang Hou [2,*]

[1] National Geomatics Center of China, Beijing 100830, China
[2] School of Geosciences and Info-Physics, Central South University, Changsha 410083, China
* Correspondence: houdongyang1986@csu.edu.cn

Abstract: Temporal intent is an important component of events. It plays an important role in collecting them from the web with focused crawlers. However, traditionally focused crawlers usually only consider factors such as topic keywords, web page content, and anchor text, ignoring the relationship between web pages and the temporal intent of events. This leads to their poor crawling performance. This paper aims to understand the temporal intent of events and apply it within focused crawlers. First, a new temporal intent identification method is proposed based on Google Trends data. The method can automatically identify the start time of an event and quantify the temporal distribution of the event. Then, a new focused event crawler with temporal intent is proposed. The crawler incorporates the start time of the event into the similarity calculation module, and a new URL (Uniform Resource Locator) priority assignment method is developed using the quantified temporal distribution of temporal intent as the independent variable of a natural exponential function. Experimental results show that our method is effective in identifying the start time of events at the month level and quantifying the temporal distribution of events. Furthermore, compared to the traditional best-first crawling method, the precision of our method improves by an average of 10.28%, and a maximum of 25.21%. These results indicate that our method performs better in retrieving relevant pages and assigning URL priority. This also illustrates the importance of the relationship between web pages and the temporal intent of events.

Keywords: event collecting; focused crawler; temporal intent; URL priority assignment

Citation: Wu, H.; Hou, D. A Focused Event Crawler with Temporal Intent. *Appl. Sci.* **2023**, *13*, 4149. https://doi.org/10.3390/app13074149

Academic Editor: Ricardo Castedo

Received: 24 February 2023
Revised: 18 March 2023
Accepted: 21 March 2023
Published: 24 March 2023

1. Introduction

With the availability of ever-increasing World Wide Web resources, there is a strong need to improve the collection of web data on events to support analysis of similar nature or man-made disasters [1,2], emergency risk assessment [3], terrorist attacks comparisons [4], and other event responses in the future. For example, it is possible to compare events or analyse trends using the information on similar events that have occurred in previous years. In general, events have an explicit or implicit temporal intent that includes when the event starts, when the event ends, and the temporal distribution of the event [5]. A simple example of this is the event "Typhoon Haiyan" with a temporal intent between the 3 and the 12 November 2013. This means that more web data about this event was published during this period. Therefore, temporal intent should be taken into account in the collection and archiving of web data on key events.

Currently, there are two ways to automatically collect and archive web data on key events: general-purpose crawler-based methods and focused crawler-based methods [6,7]. The general-purpose crawler-based methods utilize general-purpose search engines (e.g., Google and Baidu), their application programming interfaces (APIs), or universal crawlers with pre-defined queries to crawl the Internet for web data collection and archiving [8]. However, the results of these methods usually contain enormous duplicates or unrelated information [9,10], which can make the collection and archiving process tedious and time-consuming. For example, a search on Google with the pre-defined query "Typhoon Haiyan" on 3 April 2017 yields 433,000 results, which is reduced to 45,500 results when restricted to

the time period between 3 and 12 November 2013. Although the start and end times have already reduced the search results, there is still a huge amount of duplicate and irrelevant information left. This is because there is a lot of web data on other topics containing one or more keywords in the query. Therefore, a focused crawler, which intends to automatically collect as many web pages as possible that are relevant to the given topic and to keep the amount of irrelevant web pages collected to a minimum [11,12], has already been adapted to collect web data about key events [13]. These traditional researches focus on how to use web page content or hyperlink relationships to collect relevant web pages, or in what order to collect them. However, little attention is paid to the temporal intent of the event, which also plays an important role in the crawling process. For example, the start time can be used to filter out irrelevant web pages, and the temporal distribution of an event can have an impact on the priority of crawling web pages (a mathematical proof is described in Section 3.1). Thus, how to use temporal intent in a focused crawler to collect event information more efficiently remains an open problem.

In order to solve the above mentioned problem, this paper tries to understand the temporal intent of events and proposes a novel focused event crawler with temporal intent for automatically collecting web data about events. In contrast to previous approaches that only take into account factors such as event keywords, web page content, and hyperlink information, our approach additionally takes into account the temporal distribution of events while maintaining the above factors. This can improve the precision of the crawler by reducing the influence of irrelevant information. Our main contributions can be summed up as follows.

- We propose a new automated method for detecting temporal intent on events. This method uses Google Trends data to automatically and quantitatively estimate the start time and the temporal distribution of events, in contrast to previous manual judgements based on expert experience.
- We propose a new focused crawling framework that incorporates the temporal intent of events. In particular, the framework integrates the start time of the temporal intent into the process of topic representation and similarity computation, and its overall temporal distribution into the URL (Uniform Resource Locator) priority assignment. In addition, a new URL priority assignment method is proposed, in which the quantified temporal distribution is used as the independent variable of a natural exponential function.

The rest of this paper is structured as follows. Section 2 reviews related work on URL priority assignments and temporal focused crawlers. Section 3 provides a mathematical proof of the role of temporal intent in the crawling process and describes the new temporal intent identification method. The temporal intent-focused crawler is described in Section 4. Preliminary results are presented in Section 5, followed by some conclusions and future work in Section 6.

2. Related Work

Since the introduction of focused crawlers in 1999, various focused crawlers have been adapted to a number of other applications [14,15], such as geospatial web service discovery [16,17] and event web information collection [6]. However, most of the work carried out in these traditionally focused crawlers can fall into one of three categories: the representation of a given topic, relevance calculation or relevance classifier, and URL priority assignment. The first two categories have been described in detail in the previous literature and two conclusions can be drawn. The first one is that most of the traditional methods of topic representation are not able to distinguish topic keywords with temporal intent. The second one is that temporal intent is not considered in traditional relevance calculation methods. Therefore, only the work on the URL priority assignment is reviewed in this paper. It also discusses publications on some temporal focused crawlers.

2.1. URL Priority Assignment

URL priority assignment is responsible for determining the order in which URLs are crawled next [18]. An effective method of URL priority assignment can ensure that a maximum number of relevant pages will be crawled, while only a minimum number of irrelevant pages will be crawled [19]. Currently, there are four categories of URL priority assignment methods.

The first category fully exploits the relevance scores between the given topic and web page content, anchor text and their context, or URL strings, or other textual information for URL priority assignment [20]. These methods can make assigning URL priorities easier and more intuitive, but they ignore the link relationships between web pages. The second category, in which PageRank and HITS methods are often used, mainly relies on link analysis to prioritize the order of URLs to be crawled [21]. This category has the advantage of calculating the URL priority offline. However, it ignores the relevance between the web page content and the given topic, and the full network topology of the links is difficult to obtain. In addition, it is easily influenced by some noise links, such as navigation links and advertising links. Based on the above two categories, the third category is a kind of comprehensive approach. It not only assigns the URL priority according to the relevance score of the page, but also estimates the value of link analysis [22]. However, the whole network topology and noise links also affect this type of method. The last category is based on machine learning techniques, such as random forest classifier [23], semantic vector space model [24], cellular membrane optimization algorithm [25], fuzzy logic controller [26], deep learning [27], and so on. The crawling effect of this category is superior to other categories, but it needs to set more complex model parameters and requires large training sets. In addition, the influence of temporal distribution on URL priority assignment has not been considered in all four categories of methods above.

2.2. Temporal Focused Crawlers

Recently, temporal factors have been introduced into some focused crawlers [28,29]. In this paper, we refer to them as "temporal" focused crawlers. For example, Pereira et al. [30] designed a time-aware focused crawler based on temporal segmentation of web page text. In this method, URLs are divided into two categories: those within the temporal focus and the remaining ones, in which the first category has higher priority. However, the same priority is still given to URLs within the temporal focus. Farag et al. [28] and Wei et al. [6] have both developed a focused crawler based on an event model "<event content, place, time>". In the two crawlers, the start time and end time of the event were represented as independent elements and the time score was incorporated into the relevance score based on the temporal distance between the start time of the event and the publication time of the web page. However, the event model cannot represent the temporal distribution, and the priority of the URL is still assigned by the relevance scores between the given topic and web page content, anchor texts and their context, or URL strings. Furthermore, the temporal distribution has not yet been taken into account by the above mentioned temporal focused crawlers. This may affect the priority of web data collection. Therefore, our main contribution lies in the development of a new focused crawler by incorporating temporal distribution into URL priority assignment.

3. Mathematical Proof and Identification of Temporal Intent

3.1. Mathematical Proof of Temporal Intent Using Bayes Formula

In general, there will be little web information about the event before its occurrence. This means that compared to the total amount of web information about the event, its number may be negligible. It can therefore be concluded that the web information is irrelevant to the event if it is published before the event starts. After the event, there will be a series of reports on the web. Even after the event is over, there will be new web information about the event for a period of time. For example, on the anniversary of an important event, there will be some new web information on the occasion of this

special event. Therefore, the web information may or may not be relevant to the event if the publication time of the web information is equal to or greater than the start time of the event.

From a probability perspective, the above conclusions can be expressed as "given an event, the probability of web information published earlier than its start time is 0, and the probability of web information published on or later than its start time can be 0 or 1". Assuming that the start time of the event T is t_s and the publication time of a random web page is t, the above conclusion can be formally expressed as Equation (1) according to the definition of conditional probability.

$$\Pr(t/T) = \begin{cases} 0 & t < t_s \\ x & t \geq t_s \end{cases} \tag{1}$$

where $Pr(t/T)$ is the probability that a time t is relevant to an event T using the distribution of relevant web pages. The value x is an arbitrary number between 0 and 1.

$Pr(t/T)$ can be estimated by Bayes Formula, as shown in Equation (2).

$$Pr(t/T) = \frac{Pr(T/t) \times Pr(t)}{Pr(T)} = \frac{Pr(T/t) \times Pr(t)}{\sum_{t1 \in Time(WP)} Pr(T/t1) \times Pr(t1)} \tag{2}$$

where $Pr(t)$ represents the probability that time t contains a web page (relevant to T or not), $Pr(T)$ is the prior probability of finding a web page relevant to event T, and serves as a normalizing factor. A variable of WP represents a collection of web pages and $Time(WP)$ is a collection of publication times of WP. $Pr(T/t)$ is the probability that the web page published in t is relevant to the event T. For example, a random web page selected from the web pages published on 3 November 2013 has a higher probability of being relevant to the event "Typhoon Haiyan" than a random web page selected from the web pages published in November 2014.

Suppose the web page collection WP contains the number of $Num(WP)$ web pages and the number of $Num(WP, t)$ web pages published at time t, $\Pr(t)$ can be estimated by Equation (3).

$$Pr(t) = \frac{Num(WP, t)}{Num(WP)} \tag{3}$$

Suppose the collection of web pages related to the given event T is R_T. It contains the number of $Num(R_T)$ web pages and the number of $Num(R_T, t)$ web pages published at time t. Therefore, $Pr(T/t)$ can be estimated using the distribution of the web pages in R_T over time with Equation (4).

$$Pr(T/t) = \frac{Num(R_T, t)}{Num(WP, t)} \tag{4}$$

Equations (3) and (4) are then combined with (2) to obtain Equation (5).

$$Pr(t/T) = \frac{Num(R_T, t)}{Num(R_T)} \tag{5}$$

In general, we do not know the ground truth $Num(R_T)$ and $Num(R_T, t)$, but when time is infinitely large, the value of $Num(R_T)$ is fixed and invariable. Therefore, the probability $Pr(t/T)$ is proportional to the value of $Num(R_T, t)$. According to this, it is possible to draw two important roles of the temporal intent in the focused crawler.

The first role is that the start time of an event can be utilized to filter out irrelevant web pages. As discussed in the first paragraph of this section, when the web information is published earlier than the start time of the event ($t < t_s$), the value of $Num(R_T, t)$ is approximately equal to 0. When the publication time of the web information is equal to or greater than the start time of the event ($t \geq t_s$), the value of $Num(R_T, t)$ satisfies the inequality of $0 \leq Num(R_T, t) \leq Num(R_T)$. Combined with Equation (5), it can be proved

that Equation (1) is correct. Therefore, in the crawling process, all the web pages published earlier than t_s could be filtered out by comparing the start time t_s and the publication time t of a web page.

The second role is that the temporal distribution of an event can have an impact on the priority of web page crawling. According to Equation (5), if there are more relevant web pages at time $t1$ than at time $t2$, the value of $Pr(t1/T)$ will be greater than the value of $Pr(t2/T)$. This means that the URL with time $t1$ has a higher priority to be crawled. Therefore, the focused crawler can collect as many relevant web pages as possible, while minimising the amount of irrelevant web pages collected, by assigning a high priority to URLs at time t that contain more relevant web pages.

3.2. Identification of Temporal Intent by Google Trends

Temporal intent identification has already been studied in temporal information retrieval with the help of some priori data, such as query logs and time series data derived from Wikipedia [31,32]. Inspired by these works, we design a temporal intent identification method using some priori data from Google Trends data.

Google Trends data, which can be obtained from the website http://www.google.com/trends/ (accessed on 23 February 2023), represents a normalized portion value of all web searches performed worldwide for user-specified terms relative to the total number of searches conducted over a defined period [33]. The range of the Google Trends data is from 0 to 100 (see Figure 1), and the higher the value, the greater the search volume. It has been widely used in many fields, such as healthcare research [34,35], unemployment forecasting [36,37], international financial reporting standards [38], presidential election predictions [39], conservation culturomics [40], and so on. These applications have proved that Google Trends data is highly correlated with an event and its value size indirectly reflects the evolution of the event [41]. That is to say, the larger the value of the Google Trends data is, the more web information is published about the event. For this reason, we select Google Trends data as a priori data to identify the temporal intent of an event.

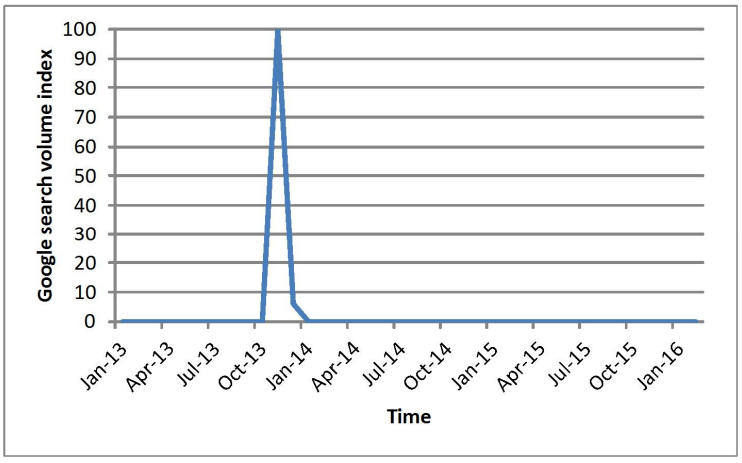


Figure 1. Google Trends data about the "Typhoon Haiyan" event.

As discussed in Section 3.1, the start time and temporal distribution in the event temporal intent play an important role in focused crawlers. Therefore, the identification of start time and temporal distribution using Google Trends data is the sole focus of this paper. The main steps of the method are described as follows.

First, the start time of an event is identified by the first fluctuation of the Google Trends data from zero to a small value. This is because, before the event, there are few web searches relative to the total number of searches about the event, which keeps its Google Trends data at zero; when the event occurs, more and more web searches are submitted, which makes its Google Trends data suddenly increase from zero. Therefore, when we get the Google Trends data of an event, the time when the non-zero value appears for the first

time is identified as the start time. For example, Figure 1 shows that the first non-zero value for the event "Typhoon Haiyan" appears in November 2013, and thus the time "November 2013" is identified as the start time of "Typhoon Haiyan".

Secondly, the temporal distribution of an event is quantized by the value of the Google Trends data. This is because its values change as the event evolves, as mentioned in the previous paragraphs. Therefore, a series of time and Google trend data are used to create quantitative temporal distributions.

4. Temporal Intent-Based Focused Crawler

The previous discussions and analyses conclude that temporal intent plays an important role in the filtering of irrelevant web pages and URL priority assignment. Therefore, the objective of this paper is to explore how temporal intent can be introduced into a focused crawler for more efficient web information collection. The conceptual framework of our temporal intent-based focused crawler is shown in Figure 2. Compared to other focused crawlers, the proposed approach involves three major tasks. The tasks in the red box are our innovative parts. The first task is the topical representation with temporal intent. This task is responsible for formalizing the topical keywords, the start time, and the temporal distribution of the given event. The second task is the relevance calculation with a start time. This task is responsible for filtering out irrelevant web pages and keeping relevant web pages using the start time. The third task is the URL priority assignment with the quantified temporal distribution. More details are given in the following subsections.

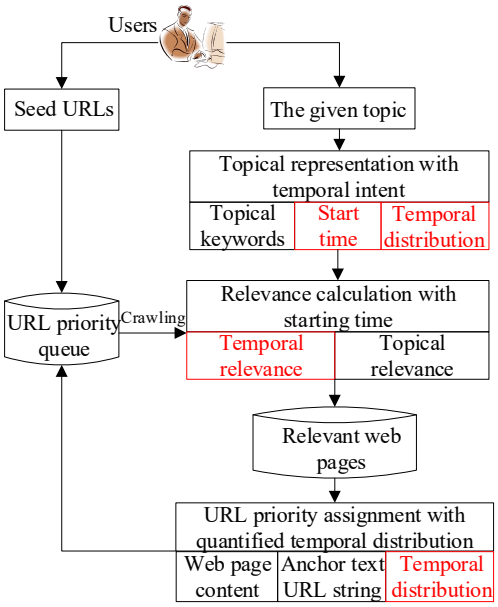


Figure 2. The conceptual framework of our focused crawler.

4.1. Topical Representation with Temporal Intent

Given a topic T about an event, it will be represented as a set of keywords in the previous works [8]. These works fail to distinguish topical keywords with temporal intent. In our proposed method, the topic T will be represented as three tuples based on the vector space model: a topical vector, the start time, and the quantized temporal distribution of the event, as shown in Equations (6)–(8).

$$T_{<>} = \langle V_{Tk}, t_{ST}, T_{TD} \rangle \tag{6}$$

$$V_{Tk} = \{(k_1, w_{Tk1}), (k_2, w_{Tk2}), \ldots, (k_n, w_{Tkn})\} \tag{7}$$

$$T_{TD} = \{< [t_{TDs1}, t_{TDe1}], \quad \lambda_1 >, \ldots, < [t_{TDsr}, t_{TDer}], \quad \lambda_r >\} \tag{8}$$

where $T_{<>}$ represents the three tuples of the event. V_{TK}, t_{ST}, T_{TD} represent the topical vector, the start time, and the quantized temporal distribution of the event, respectively. The topical vector V_{TK} consists of a series of keywords $k_1, k_2, \ldots, k_n$ and their weights $w_{Tk1}, w_{Tk2}, \ldots, w_{Tkn}$, which are calculated by the normalized term frequency with a predefined corpus in this study [42]. The quantized temporal distribution T_{TD} is composed by a set of time intervals and their quantized values λ_r, which denote the relative volume of search queries conducted through Google Trends data. The t_{TDsr} and t_{TDer} are the start and end time of the time interval.

In the crawling process, a web page D will be represented as two tuples $D_{<>}$, as shown in Equations (9) and (10). One tuple is a vector V_{Dk} composed by keywords $k_1, k_2, \ldots, k_n$ and their weights $w_{Dk1}, w_{Dk2}, \ldots, w_{Dkn}$, which are the frequencies of occurrence of keywords $k_1, k_2, \ldots, k_n$ in the web page D. The other is the publication time t_{PT} of the web page.

$$D = \langle V_{Dk}, t_{PT} \rangle \tag{9}$$

$$V_{Dk} = \{(k_1, w_{Dk1}), (k_2, w_{Dk2}), \ldots, (k_s, w_{Dks})\} \tag{10}$$

4.2. Relevance Calculation with Start Time

The start time can be used as an indicator to judge whether a web page is relevant to the given event, as discussed in Section 3.1. However, only topical keywords are usually used in traditional relevance calculation methods generally only use topical keywords. Therefore, the relevance between a given event and a web page is calculated from the hierarchy of the temporal and topical keywords in this study. In the hierarchical method, there are two steps to calculate and judge the topic's relevance.

First, the temporal relevance $sim(t_{PT}, t_{ST})$ is calculated by comparing the publication time of a web page with the start time of the given event using Equation (11).

$$sim(t_{PT}, t_{ST}) = \begin{cases} 0 & t_{PT} < t_{STs} \\ 1 & t_{PT} \geq t_{STs} \end{cases} \tag{11}$$

where variables are the same as Equations (6) and (9). If the publication time t_{pt} of the web page is earlier than the start time of the event, the value of temporal relevance is zero and the web page is abandoned. Otherwise, the value of temporal relevance is one, which indicates that the web page may be relevant to the given event.

Secondly, the topical relevance $sim(V_{Dk}, V_{Tk})$ is still calculated by using the cosine formula (as shown in Equation (12)) if the publication time t_{pt} of the web page is later than the start time of the event.

$$sim(V_{Dk}, V_{Tk}) = \frac{\sum_{i=1}^{s} w_{Tki} \times w_{Dki}}{\sqrt{\sum_{i=1}^{s} w_{Tki}^2 \times \sum_{i=1}^{s} w_{Dki}^2}} \tag{12}$$

where variables are the same as Equations (7) and (10). If the relevance $sim(V_{Dk}, V_{Tk})$ is greater than or equal to the specific threshold, it means that the web page is relevant to the given event and the focused crawler will store the web page in a web page repository. Otherwise, the web page will be discarded.

4.3. URL Priority Assignment with Quantified Temporal Distribution

As mentioned in Section 3, the priority of a URL published at time t is proportional to the number of relevant web pages, which can be represented by the Google search volume index. Therefore, a natural exponential function, which is an increasing function and is commonly used in time series analysis [43], is constructed and combined into the URL

priority assignment with the Google search volume index. The final equations of the new method are shown in Equations (13) and (14).

$$P(URL) = \theta \times sim(V_{Dk}, V_{Tk}) + \gamma \times sim_{anchor} \tag{13}$$

$$P_T(URLL) = \begin{cases} P(URL) & P(URL) < \delta \\ exp(\frac{\lambda_i}{max\{\lambda_1,\ldots\lambda_r\}}) \times P(URL) & P(URL) \geq \delta \end{cases} \tag{14}$$

where $P(URL)$ and $P_T(URL)$ represent the URL priority computed by the traditional and proposed methods, respectively. Variable $sim(V_{Dk}, V_{Tk})$ is the same as Equation (12). Variable sim_{anchor} denotes the anchor relevance. Variables θ and γ are weighted factors, and they satisfy $\theta + \gamma = 1$ ($\theta \geq 0$ *and* $\gamma \geq 0$). In Equation (13), the anchor text is a direct description of the URL and is more important for the URL priority than the content of the parent page. Therefore, variables θ and γ are set to 0.4 and 0.6, respectively. The $exp()$ represents the natural exponential function. The variable λ_i is the same as Equation (8). The $max\{\ldots\}$ denotes the maximum value of λ_i. The variable δ is a priority reduction factor and is set as 0.4 through many experiments. It is used to ensure that the priority of the URL corresponding to a relevant web page is raised and meanwhile the priority of the URL corresponding to an irrelevant web page remains unchanged.

In Equation (13), the anchor relevance sim_{anchor} is also calculated by the cosine formula using the anchor vector V_{AK} and topical vector V_{TK}, as shown in Equations (15) and (16).

$$V_{AK} = \{(k_1, w_{Ak1}), (k_2, w_{Ak2}), \ldots, (k_n, w_{Akn})\} \tag{15}$$

$$sim(V_{Ak}, V_{Tk}) = \frac{\sum_{i=1}^{s} w_{Tki} \times w_{Aki}}{\sqrt{\sum_{i=1}^{s} w_{Tki}^2 \times \sum_{i=1}^{s} w_{Aki}^2}} \tag{16}$$

where w_{Aki} represents the weight of the keyword k_i and is calculated by the frequency of occurrence of keywords k_i in the anchor text. Other variables are the same as in Equation (7).

In Equation (14), the value of the variable λ_i is related to the publication time of the corresponding URL. In this paper, two ways are used to extract the publication time. The first way is to use regular expressions to extract the time from the URL, which contains a time expression. For example, the time of 5 September 2008 can be extracted from the URL of 'http://news.sohu.com/20080905/n259388056.shtml' (accessed on 23 February 2023). If the URL does not contain a time expression, then its time of publication will be assumed to be the same as the time of publication of its parent web page. In fact, its publication time is equal to or slightly earlier than the publication time of its parent web page. However, each Google search volume index corresponds to a larger time interval, which means that two publication times with a small gap are very likely to have the same Google search volume index. Therefore, the hypothesis of the second way is reasonable.

5. Experiments and Discussion

5.1. Experimental Setup

To verify the effectiveness of our approach, we first conducted experiments on event temporal intent identification, followed by experiments on the effectiveness of our focused crawler. In the experiments, the above method was implemented using the C# language. The experiments were conducted in an environment with an Intel Pentium 4 CPU 3.20 GHz, 1 GB RAM, and 6 M bandwidth.

5.1.1. Effectiveness Metrics

There are two basic effectiveness metrics for focused crawlers [8]. The first is the precision, which denotes the fraction of relevant web pages in the crawled web pages. The higher the precision value, the better the crawler's ability to prioritize URLs. The second

is the recall, which represents the fraction of the crawled relevant web pages in the total relevant web pages. The higher the recall value, the better the crawler's ability to retrieve relevant web pages. In this paper, only the precision is chosen as our effectiveness metric, because the number of relevant web pages in the whole web is unknown, and this paper mainly investigates whether the temporal intent can help the crawler to find relevant web pages quickly. It can be calculated with Equation (17).

$$p = \frac{CR}{TC} \tag{17}$$

where variables p represents the precision. The variable CR denotes the number of relevant pages crawled. The variables TC indicate the total number of crawled pages.

5.1.2. Data Preparation

According to the definition of a focused crawler, topical keywords must be identified prior to the crawling experiment. In this paper, a time-related event "Typhoon Haiyan" was chosen to test the effectiveness of our focused crawler. Specifically, "Typhoon Haiyan" was submitted as a query to Baidu News Search to identify the topical keywords. Based on the search results, 100 pages related to "Typhoon Haiyan" were manually selected. The content of the 100 relevant web pages was then segmented into different keywords using Pangu Chinese word segmentation, and these keywords were counted by word frequency. The keywords were then ranked according to their frequency, and 12 keywords with high frequency and practical importance were selected as topical keywords. It should be noted that synonyms were combined in this process, e.g., "typhoon" and "hurricane". The weight of the topical keywords was then calculated using the document-inverse document word frequency method. The final set of topical keywords and their weight for the event "Typhoon Haiyan" are shown in Equation (18).

$$V_{Tk} = \left\{ \begin{array}{l} (typhoon, 1), (Haiyan, 0.92), (Philippines, 0.57), \\ (Guangxi, 0.45), (rescue, 0.32), (Sanya, 0.26), \\ (heavyrain, 0.24), (death, 0.2), (Hainan, 0.2), \\ (Vietnam, 0.2), (disaster, 0.19), (landfall, 0.16) \end{array} \right\} \tag{18}$$

The initial URLs for the crawling experiment were selected manually and automatically based on search engines. First, 12 web pages about Typhoon Haiyan from major portals were manually selected. Then, "Typhoon Haiyan" was submitted as a query to the Baidu News search engine and 360 News search engines, and 100 non-repeating URLs were obtained. Finally, 112 URLs were used as the initial URLs for this experiment.

5.2. Experiment 1: Temporal Intent Identification

This section presents an experiment that demonstrates the ability of Google Trends to identify temporal intent. The experiment takes "WenChuan earthquake", "YuShu earthquake", "Typhoon HaiYan", "Haiti earthquake", and "Indian Ocean Tsunami" as examples. Their Google Trends data is shown in Figure 3.

As mentioned in Section 3.2, the start times of these events are identified according to the first fluctuation of the Google Trends data from zero to a small value. Table 1 shows the comparison between the detected start time and the actual start time. It can be seen from Table 1 that all the identified start times are consistent with their actual start times at the month level. This shows that the Google Trends data is closely related to the events. It also shows that the method proposed in this paper for identifying start times based on Google Trends data is effective.

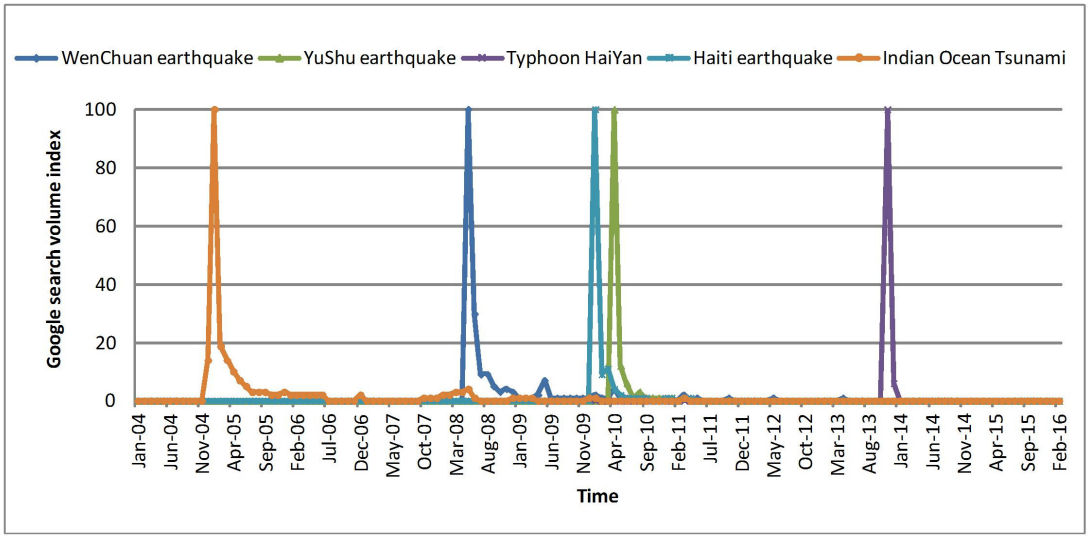


Figure 3. Examples of Google Trends.

Table 1. Comparisons between the detected and actual start time.

Event Name	Time Interval When the Initial Search Volume Index Is Zero	Time When the Initial Search Volume Index Is Greater than Zero	Detected Start Time	Actual Start Time
WenChuan earthquake	[2004-01, 2008-04]	2008-05	2008-05	2008-05-21
YuShu earthquake	[2004-01, 2010-03]	2010-04	2010-04	2010-04-14
Typhoon HaiYan	[2004-01, 2013-10]	2013-11	2013-11	2013-11-03
Haiti earthquake	[2004-01, 2009-12]	2010-01	2010-01	2010-01-12
Indian Ocean Tsunami	[2004-01, 2004-11]	2004-12	2004-12	2004-12-26

Furthermore, the values of the Google Trends data are also a reflection of the temporal distribution of these events in Figure 3. For example, the event "WenChuan earthquake" was of high concern to users from May to December 2008 and was renewed in May 2009. This is consistent with its Google Trends data in Figure 3. This shows that it is reasonable to use Google Trends data to quantify the temporal distribution of the event.

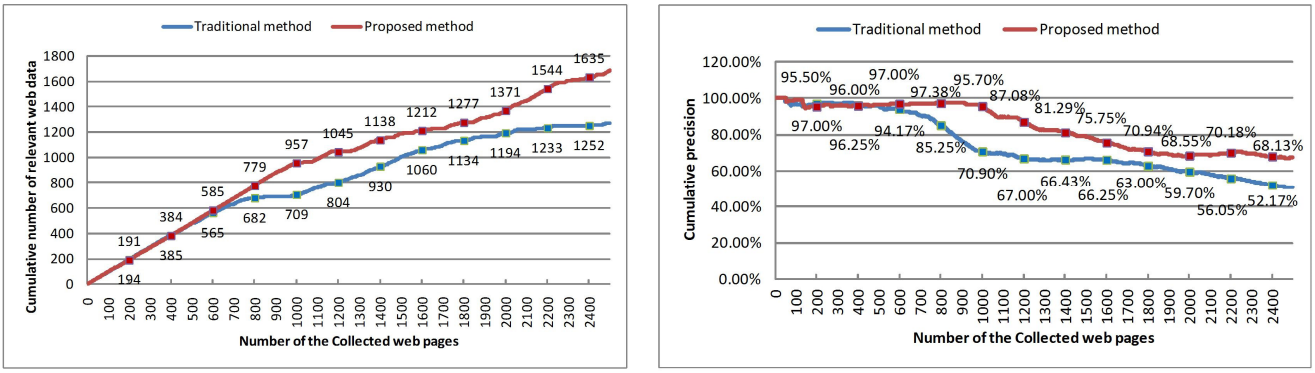
Based on the above analysis, the start time and quantized temporal distribution of the event "Typhoon Haiyan" in the following experiments are shown as Equations (19) and (20), whose variables are the same as in Section 4.1.

$$T_{ST} = [\text{2013-11-01}, \infty] \tag{19}$$

$$T_{TD} = \{< [\text{2013-11-01}, \text{2013-11-30}, 100 >, < [\text{2013-12-01}, \text{2013-12-31}], 6 >\} \tag{20}$$

5.3. Experiment 2: Effectiveness of Our Focused Crawler

This experiment uses the mainstream best-first crawling strategy-based focused crawler as a benchmark for comparison. The best-first crawling strategy mainly uses the cosine similarity between topic keywords and web content as the similarity calculation and assigns URL priority using web content similarity and anchor text similarity. Our temporal intent focused crawler is mainly based on the formula in Section 4. In the experiments, the two methods use the same initial URL, topical keywords, web page request method, and web page parsing method, while the differences are in the topic representation method, topic relevance calculation strategy, and URL priority assignment method. Finally, 2500 web pages were crawled by each of the two methods to evaluate the crawling performance. The experimental results for the two crawling methods are shown in Figure 4.



(**a**) (**b**)

Figure 4. Crawling results of the two methods. (**a**) is the cumulative number of relevant web pages. (**b**) is cumulative precision.

Figure 4a depicts the cumulative number of relevant web pages. In other words, for any number n of crawled web pages, the total number of the crawl up to that point is reported. From Figure 4a, it can be seen that in the initial stage of crawling (the first 600 web pages), the number of relevant web pages crawled by the two methods is almost the same. This is mainly because they use the same initial URL, and the web pages obtained in the initial stage are chained from the initial URL. The number of relevant pages crawled by our method was significantly higher than that crawled by the traditional best-first method after crawling 600 pages. For example, our focused crawler obtains 1138 relevant web pages while the traditional one obtains 930 relevant web pages when the two crawlers both collected 1400 pages. Therefore, it is intuitive that the crawling effect of the focused crawler in this paper is superior to that of the traditional method.

In order to perform a quantitative comparative analysis, the precision metric is computed according to Equation (17). Figure 4b depicts the cumulative precision of the two crawlers. As can be seen in Figure 4b, the difference in precision between the two methods is small for the first 600 crawled web pages, with a difference of 0.1% in their cumulative average precision. After that, the precisions of our method are consistently higher than those of the traditional best-first method, with an average improvement of 10.28%. The greatest improvement in precision is achieved when a total of 974 web pages are crawled, with an increase of 25.21%. The results indicate that the ability of our method to find pages related to a topic event is better than the traditional method.

There are two main reasons why our method is able to achieve such excellent results. First, our method can identify the temporal intent of the event (i.e., the start time and the quantified temporal distribution) and can filter out pages published before the event using a relevance calculation strategy based on the start time. More importantly, a new URL priority assignment method is designed based on the quantified temporal distribution, which can prioritize the crawling of pages published at the time of the event. In contrast, traditional best-first methods still focus only on the keywords of the event and the content of the web page, while ignoring the relationship between the web page and the temporal intent of the event. This leads to traditional methods of crawling many irrelevant web pages before and after the event occurred.

6. Conclusions and Future Work

The collection of web information about the event can be seriously affected by the temporal intent of the event. However, the temporal intent is still ignored by current focused crawling methods. In this paper, we propose a new temporal intent identification method and a novel focused event crawler with temporal intent. In the new temporal intent identification method, the Google Trends data is used to automatically identify the start time of events and quantify the temporal distribution of events. In the new focused event crawler, the identified start time is introduced into relevance calculation to filter out irrelevant web

pages published before the event occurred. Furthermore, the crawler constructs a natural exponential function with the quantified temporal distribution to change the crawling priority. The experimental results show that our method can effectively identify the start time of events at the monthly level and quantify the relationships between the temporal distribution of events and the number of web pages. In addition, the crawling experimental results show an average improvement of 10.28% and a maximum improvement of 25.21% in the crawling precision of our method in comparison with the traditional best-first crawling method.

Although our method has improved significantly in terms of crawling effectiveness, there is still room for improvement. Currently, the temporal intent of events is only identified based on Google Trends data. This data is limited by the geographic nature of Google users and therefore does not cover all events. As a result, we may not be able to identify the temporal intent of certain events. In the future, we will add more a priori data (e.g., the Baidu index). We will also create a feedback mechanism using the crawled web data. In addition, spatial location is an important component of events. It has been shown to be effective in focused crawlers [8]. However, the spatial semantics of events are not considered in the approach of this paper. Therefore, we will try to add a spatial similarity calculation module to the temporal intent focused crawler. In addition, we are currently still optimizing the traditional crawling method. In the future, we will also consider combining some optimization algorithms from other domains (e.g., Runge-Kutta optimization algorithm and meta-heuristic algorithm [44–46]) to further improve the crawling effect.

Author Contributions: Conceptualization, H.W. and D.H.; methodology, H.W. and D.H.; software, H.W. and D.H.; validation, H.W. and D.H.; writing—original draft preparation, H.W. and D.H.; project administration, D.H.; funding acquisition, D.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded in part by Hunan Provincial Natural Science Foundation of China, grant number 2021JJ40721 and in part by Yunnan Fundamental Research Projects, grant number 202001AS070032.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare no conflict of interest. The authors would also like to thank the anonymous reviewers for their comments to improve this paper.

References

1. Franceschini, R.; Rosi, A.; Catani, F.; Casagli, N. Exploring a landslide inventory created by automated web data mining: The case of Italy. *Landslides* **2022**, *19*, 841–853. [CrossRef]
2. Sufi, F.K.; Khalil, I. Automated Disaster Monitoring from Social Media Posts Using AI-Based Location Intelligence and Sentiment Analysis. *IEEE Trans. Comput. Social Syst.* 2022; *early access*. [CrossRef]
3. Huang, X.; Jin, H.D.; Zhang, Y. Risk assessment of earthquake network public opinion based on global search BP neural network. *PLoS ONE* **2019**, *14*, e0212839. [CrossRef] [PubMed]
4. Amiresmaili, M.; Talebian, A.; Miraki, S. Pre-hospital emergency response to terrorist attacks: A scoping review. *Hong Kong J. Emerg. Med.* **2022**, *29*, 56–62. [CrossRef]
5. Campos, R.; Dias, G.; Jorge, A.M.; Jatowt, A. Survey of temporal information retrieval and related applications. *ACM Comput. Surv. (CSUR)* **2014**, *47*, 15. [CrossRef]
6. Wei, X.; Hu, H.; Zeng, D.D.; Wu, W. Emergency Event Web Information Acquisition using Crowd Web Sensors. *Wirel. Pers. Commun.* **2017**, *95*, 2393–2411. [CrossRef]
7. Neelakandan, S.; Arun, A.; Bhukya, R.R.; Hardas, B.M.; Kumar, T.C.A.; Ashok, M. An Automated Word Embedding with Parameter Tuned Model for Web Crawling. *Intell. Autom. Soft Comput.* **2022**, *32*, 1617–1632. [CrossRef]
8. Hou, D.; Wu, H.; Chen, J.; Li, R. A Focused Crawler for Borderlands Situation Information with Geographical Properties of Place Names. *Sustainability* **2014**, *6*, 6529–6552. [CrossRef]

9. Shi, Q.; Shi, Z.; Xiao, Y. VSEC: A Vertical Search Engine for E-commerce. In *Recent Progress in Data Engineering and Internet Technology*; Springer: Berlin/Heidelberg, Germany, 2012; pp. 57–63.
10. Lupiani-Ruiz, E.; García-Manotas, I.; Valencia-García, R.; García-Sánchez, F.; Castellanos-Nieves, D.; Fernández-Breis, J.T.; Camón-Herrero, J.B. Financial news semantic search engine. *Expert Syst. Appl.* **2011**, *38*, 15565–15572. [CrossRef]
11. Liu, J.F.; Li, X.; Zhang, Q.S.; Zhong, G. A novel focused crawler combining Web space evolution and domain ontology. *Knowl.-Based Syst.* **2022**, *243*, 108495. [CrossRef]
12. Tchakounte, F.; Ngnintedem, J.C.T.; Damakoa, I.; Ahmadou, F.; Fotso, F.A.K. Crawl-shing: A focused crawler for fetching phishing contents based on graph isomorphism. *J. King Saud Univ.-Comput. Inf. Sci.* **2022**, *34*, 8888–8898. [CrossRef]
13. Wang, J.; Dang, D.; Zhou, P.; Wang, H.; Jiang, X.; Huang, S. Crawling Strategy Based on Domain Ontology of Emergency Plans. In Proceedings of the 2013 the International Conference on Education Technology and Information System (ICETIS 2013), Sanya, China, 21–22 June 2013.
14. Chuang, H.M.; Chang, C.H.; Kao, T.Y.; Cheng, C.T.; Huang, Y.Y.; Cheong, K.P. Enabling maps/location searches on mobile devices: Constructing a POI database via focused crawling and information extraction. *Int. J. Geog. Inf. Sci.* **2016**, *30*, 1405–1425. [CrossRef]
15. Jung, J.J. Towards open decision support systems based on semantic focused crawling. *Expert Syst. Appl.* **2009**, *36*, 3914–3922. [CrossRef]
16. Hou, D.; Chen, J.; Wu, H. Discovering Land Cover Web Map Services from the Deep Web with JavaScript Invocation Rules. *ISPRS Int. J. Geo-Inf.* **2016**, *5*, 105. [CrossRef]
17. da Silva, A.S.; Lisboa-Filho, J. A Focused Crawler for Web Feature Service and Web Map Service Discovering. In Proceedings of the Web and Wireless Geographical Information Systems: 18th International Symposium, W2GIS 2020, Wuhan, China, 13–14 November 2020; p. 111.
18. Capuano, A.; Rinaldi, A.M.; Russo, C. An ontology-driven multimedia focused crawler based on linked open data and deep learning techniques. *Multimed. Tools Appl.* **2020**, *79*, 7577–7598. [CrossRef]
19. Dang, T.K.N.; Bucur, D.; Atil, B.; Pitel, G.; Ruis, F.; Kadkhodaei, H.; Litvak, N. Look back, look around: A systematic analysis of effective predictors for new outlinks in focused Web crawling. *Knowl.-Based Syst.* **2023**, *260*, 110126. [CrossRef]
20. Talvensaari, T.; Pirkola, A.; Järvelin, K.; Juhola, M.; Laurikkala, J. Focused web crawling in the acquisition of comparable corpora. *Inf. Retr.* **2008**, *11*, 427–445. [CrossRef]
21. Alam, M.H.; Ha, J.; Lee, S. Novel approaches to crawling important pages early. *Knowl. Inf. Syst.* **2012**, *33*, 707–734. [CrossRef]
22. Boukadil, K.; Rekik, M.; Rekik, M.; Ben-Abdallah, H. FC4CD: A new SOA-based Focused Crawler for Cloud service Discovery. *Computing* **2018**, *100*, 1081–1107. [CrossRef]
23. Rajiv, S.; Navaneethan, C. A Supervised Learning-Based Approach for Focused Web Crawling for IoMT Using Global Co-Occurrence Matrix. *Expert Syst.* 2022; *early access*. [CrossRef]
24. Liu, W.; Gan, Z.; Xi, T.; Du, Y.; Wu, J.; He, Y.; Jiang, P.; Liu, X.; Lai, X. A semantic and intelligent focused crawler based on semantic vector space model and membrane computing optimization algorithm. *Appl. Intell.* **2022**, *53*, 7390–7407. [CrossRef]
25. Liu, W.J.; Du, Y.J. A novel focused crawler based on cell-like membrane computing optimization algorithm. *Neurocomputing* **2014**, *123*, 266–280. [CrossRef]
26. Sharma, R.; Bhatia, R.; Garg, S.; Aujla, G.S.; Mann, R.S. Fuzzy Based Efficient Mechanism for URL Assignment in Dynamic Web Crawler. In *Advanced Informatics for Computing Research: First International Conference, ICAICR 2017, Jalandhar, India, 17–18 March 2017, Revised Selected Papers*; Singh, D., Raman, B., Luhach, A.K., Lingras, P., Eds.; Springer: Singapore, 2017; pp. 3–17.
27. Shrivastava, G.K.; Pateriya, R.K.; Kaushik, P. An efficient focused crawler using LSTM-CNN based deep learning. *Int. J. Syst. Assur. Eng. Manag.* **2023**, *14*, 391–407. [CrossRef]
28. Farag, M.M.; Lee, S.; Fox, E.A. Focused crawler for events. *Int. J. Digit. Libr.* **2018**, *19*, 3–19. [CrossRef]
29. Klein, M.; Balakireva, L.; Van de Sompel, H. Focused crawl of web archives to build event collections. In Proceedings of the 10th ACM Conference on Web Science, Amsterdam, The Netherlands, 27–30 May 2018; pp. 333–342.
30. Pereira, P.; Macedo, J.; Craveiro, O.; Madeira, H. Time-Aware Focused Web Crawling. In *Advances in Information Retrieval*; Springer: Berlin/Heidelberg, Germany, 2014; pp. 534–539.
31. Zhao, Y.; Hauff, C. Temporal Query Intent Disambiguation using Time-Series Data. In Proceedings of the 39th International ACM SIGIR Conference on Research and Development in Information Retrieval, Pisa, Italy, 17–21 July 2016; pp. 1017–1020.
32. Campos, R.; Jorge Alípio, M.; Dias, G. Using Web Snippets and Web Query-logs to Measure Implicit Temporal Intents in Queries. In Proceedings of the 2nd Workshop on Query Representation and Understanding of the 34th ACM Annual SIGIR Conference (SIGIR 2011), Beijing, China, 24–28 July 2011. 4p.
33. Jun, S.P.; Yoo, H.S.; Choi, S. Ten years of research change using Google Trends: From the perspective of big data utilizations and applications. *Technol. Forecast. Soc. Chang.* **2018**, *130*, 69–87. [CrossRef]
34. Xu, Y.W.; Margolin, D. Collective Information Seeking during a Health Crisis: Predictors of Google Trends during COVID-19. *Health Commun.* 2023; *early access*. [CrossRef]
35. Arora, V.S.; McKee, M.; Stuckler, D. Google Trends: Opportunities and limitations in health and health policy research. *Health Policy* **2019**, *123*, 338–341. [CrossRef]
36. Simionescu, M.; Cifuentes-Faura, J. Can unemployment forecasts based on Google Trends help government design better policies? An investigation based on Spain and Portugal. *J. Policy Model.* **2022**, *44*, 1–21. [CrossRef]

37. Simionescu, M.; Cifuentes-Faura, J. Forecasting National and Regional Youth Unemployment in Spain Using Google Trends. *Soc. Indic. Res.* **2022**, *164*, 1187–1216. [CrossRef]
38. Zhang, Y. Using Google Trends to Track the Global Interest in International Financial Reporting Standards: Evidence from Big Data. *Intell. Syst. Account. Financ. Manag.* 2023; *early access*. [CrossRef]
39. Vergara-Perucich, F. Assessing the Accuracy of Google Trends for Predicting Presidential Elections: The Case of Chile, 2006–2021. *Data* **2022**, *7*, 143. [CrossRef]
40. Correia, R.A.; Ladle, R.; Jaric, I.; Malhado, A.C.M.; Mittermeier, J.C.; Roll, U.; Soriano-Redondo, A.; Verissimo, D.; Fink, C.; Hausmann, A.; et al. Digital data sources and methods for conservation culturomics. *Conserv. Biol.* **2021**, *35*, 398–411. [CrossRef] [PubMed]
41. Chen, T.; Lin, J. Comparative Analysis of Temporal-Spatial Evolution of Online Public Opinion Based on Search Engine Attention:Cases of Google Trends and Baidu Index. *J. Intell.* **2013**, *32*, 7–10+16.
42. Liu, B. *Web Data Mining: Exploring Hyperlinks, Contents, and Usage Data*, 2nd ed.; Springer: New York, NY, USA, 2010; pp. 217–218.
43. Li, X.; Liu, B.; Philip, S.Y. Time sensitive ranking with application to publication search. In *Link Mining: Models, Algorithms, and Applications*; Springer: Berlin/Heidelberg, Germany, 2010; pp. 187–209.
44. Devi, R.M.; Premkumar, M.; Jangir, P.; Elkotb, M.A.; Elavarasan, R.M.; Nisar, K.S. IRKO: An Improved Runge-Kutta Optimization Algorithm for Global Optimization Problems. *CMC-Comput. Mater. Contin.* **2022**, *70*, 4803–4827. [CrossRef]
45. Gupta, D.; Dhar, A.R.; Roy, S.S. A partition cum unification based genetic- firefly algorithm for single objective optimization. *Sadhana* **2021**, *46*, 121. [CrossRef]
46. Ghasemi, M.; Akbari, M.A.; Jun, C.Y.; Bateni, S.M.; Zare, M.; Zahedi, A.; Pai, H.T.; Band, S.S.; Moslehpour, M.; Chau, K.W. Circulatory System Based Optimization (CSBO): An expert multilevel biologically inspired meta-heuristic algorithm. *Eng. Appl. Comput. Fluid Mech.* **2022**, *16*, 1483–1525. [CrossRef]