Framework for Evaluating Credibility of External Links in Wikipedia

Imran Latif, Syed Waqar Jaffry, Shahzad Sarwar, Laeeq Aslam, Muhammad Murtaza Yousaf

***Abstract* — On the advent of Web 2.0, web users have graduated from mere information-consumers and have become information-producers. Wikipedia is one of the paramount examples of this phenomenon. Open collaborative editing model of Wikipedia allows anyone to contribute information, from anywhere. Hence, general public and particularly researchers are skeptical about the information available at Wikipedia. The whole data set of Wikipedia that ranges from Wikipedia content to editors' communication is publically available and open to use. Using this information, it seems practical to design models and frameworks to measure authenticity of Wikipedia content. In order to measure authenticity of Wikipedia information, external links play an important role. The presence of external sources or links on a web page to other sites can increase credibility of information as it allows visitors to cross-check information at external sites. However, there should be some mechanism to validate these external sources. In this work, an External Link Verification framework has been proposed and evaluated on External Links of Wikipedia articles. The proposed framework could be used to compute credibility of an external link of any web page.**

***Index Terms* — External Links, Sources Verification, Wikipedia.**

I. INTRODUCTION

In online contents, usually references are provided for the sake of information verifiability and credibility but in the literature a framework to check the credibility of these references is missing. The same problem is also faced by the largest web-based encyclopedia that is Wikipedia. There is no framework for the quality assessment of external links referred in an article [1]. So far, the impact of reference credibility on the article's overall quality and credibility has not been examined on account of guidelines for referencing by the Wikipedia [1]. It is assumed, the guidelines ensure the additive value of external links to the article. The fundamental purpose behind this work is to design a framework that ensures the credibility of external links, referred in Wikipedia articles. So that based on web sources, overall credibility of the information provided by the Wikipedia as an encyclopedia can be gauged.

II. BACKGROUND AND LITERATURE REVIEW

The phenomenal growth in the Wikipedia content actually makes it the largest encyclopedia. Absence of the explicit user/peer review system to the process of content verification of the Wikipedia makes it extremely difficult for both the administrators and the editors to validate and coordinate the content on regular basis. Such coordination is required to ensure that external links are credible that provide relevant and authentic information. Also, the article merit can be calculated through the automatic validation of the credibility of external links.

The research community is skeptical about the quality of Wikipedia article. There are several studies that focused on numerous quality assessment methods. Usually, these methods depend on the analysis of internal attributes of the article for instance number of edits [2], length of article [3], contextual elements [4], formality of article language [5], article linkage within the Wikipedia article graph [6], and factual accuracy [7]. It has been observed that a systematic processing is not in place to verify credibility of the external links referred in a Wikipedia article. In this paper, we investigate the role of external links to verify the accuracy of information provided in an article, which is indeed an essential component of the overall Wikipedia trust evaluation framework.

One of the main motives for this study is the possibility of quantification of the external links credibility and its usefulness towards improving merit of the article. The statistical distribution of different links provided in Wikipedia articles and interesting facts like most well-known domain names among external sources of an article can be identified [8]. Subsequently, these facts can be used to foster the process of calculating credibility of the article. Also, the correlation between fractions of article's external links with their content size can be examined.

A. Wikipedia External link Guidelines

Wikipedia editing guidelines provides instructions that links provided in the article should point to the information that is not described within the article. So this work can verify the appropriateness of external links regarding the article. The decay of Wikipedia external links can also be quantified in order to assist the administrators and editors. It can also assist to determine the amount of repair required regarding the dead links with in articles.

A web page that has not any web redirects or no longer available to the web user can be pointed as a Deadlink page [9]. The other way to identify the dead or irrelevant links is to examine such links that have no relevance with the original article. Proportion of dead links to estimates the decay of an article's external pages can also be calculated. One can measure appropriateness of the linked pages with the article's subject and calculate the relatedness of linked resources with the article's subject.

B. External Links' Distribution

Wikipedia External links can be easily identified as they are present at distinct sections of the article's body. Using

Imran Latif is from IBM Global Business Services, Pakistan, Syed Waqar Jaffry, Shahzad Sarwar, Laeeq Aslam and Muhammad Murtaza Yousaf are with Punjab University College of Information Technology (PUCIT), University of the Punjab, Lahore, Pakistan. Email: imisweet1@gmail.com, swjaffry@pucit.edu.pk,s.sarwar@pucit.edu.pk, laeeq.aslam@pucit.edu.pk, murtaza@pucit.edu.pk. Manuscript received Feb 25, 2017 revised on April 05, 2017. Accepted on May 27, 2017.

 ISSN – 1999-4974

Wikipedia dumps [10] one can easily identified the value of external links for some article. To capture the correlation and appropriateness of references with the Wikipedia article, usually MediaWiki API [11] is used.

One can address this problem by identifying the non-wiki sources for each article. It is not straightforward job as there are several ways to link an external resource with an article. One can handle it by merging all identified links into a single file, and remove the bullet and space characters that are present in start of every non-wiki link. Later on these links can be organized in form of clusters according to their syntax. Following are some expected emerging clusters together during the processing of external links:

- URLs that are placed in square brackets, followed by its hyper linked label containing a serial number
- Hyperlink URL names
- URLs that are placed in square brackets, followed by a space character and text used to label the hyperlink

The fundamental task is to parse URLs of the external resources and calculate their distribution in the Wikipedia collection by identifying their domain names. To compute the distribution of the external resource's domain name within the Wikipedia one can simply count the number of articles that contains the same external link of the identified domain. Other important aspect that one needs to examine is the correlation between the amount of external links and the articles' length. To achieve this one needs to perform the following steps:

- Parse the Wikipedia dump's XML file to get pure text related to article or get the text using MediaWiki API.
- Then based on the article's text quantify the article's length using count of characters that it contains.
- Make a vector of paired numerical values that represents the amount of its external resources and article's length.
- This vector can visualize the correlation between non-wiki linked resources and the article's text size.

There are some observations as already presented in [16] that out of the 3,290,179 English Wikipedia articles around 23.91% (786,857) articles provide no links to external resources. While the remaining 76.09% (2,503,322) articles contain 13,355,687 external links. This results into the Wikipedia ratio of article to external links around 4.06 with median around 2. Furthermore around 7.30% of the articles point to more than 100 external links, and 68.79% of the articles contain up to 10 external links.

C. Dead External Links

A web page that is not accessible over the web is known as a dead page. Using the fraction of dead pages amongst non-wiki linked sources one can identify the decay in the Wikipedia external links. One can determine the dead pages easily using following steps:

- Check the web URL that either it fails on URL parsing
- Check the web URL that either it fails on resolution of its host address
- Check that either it returns the HTTP error
- Check for the redirect pages not returning the 3xx series
- Check that landing page return OK code or an error 404 HTTP

According to a proposal presented in [12] in case of the 404 error one can securely assume that the page is dead but in case of the HTTP OK code there are two possible ways. One can only consider the page is dead if it fulfills the following conditions:

- If an original URL page is redirecting to the home page of the host or non-existent randomly generated URL (system generated fake URL made by appending some random charters to original URL) is also redirecting to the home page.
- The external link redirects to another URL and that also redirects, so that its retrieval enters in a loop.
- The reference link or randomly generated URL refers to a page that is nearly identical calculated via shingling [13]. Shingling is text similarity measurement technique in which shingles (unique n-grams) of the source text are compared with the target text [13].

The page is considered live only if it fulfills the following conditions:

- It is website's root.
- It redirects to a URL in same host directory.
- Failure HTTP code is return for non-existing randomly generated URL for the same host directory.

An elegant algorithm presented in [15] covers the above situation and also encapsulate the interpretation of it. This algorithm identifies the dead pages. One needs to run this algorithm on an extracted Wikipedia external link for verification. Usually when accessing web pages on the web if host does not have the desired webpage then the web server of the host returns 404 HTTP code. However it is also found that some commodity web servers' returns nothing in the same scenario. Instead they return an OK code (200) with some error message on a customized webpage or even redirect to some other page. Such pages which does not exist and behave as stated above are known as "soft-404 pages".

In order to find the proportion of dead links in Wikipedia external resources, some studies are performed. For example in [14] it is claimed that external links of nearly 2 million articles contain no dead links at all while around 18.34% of the examined links are dead which returned hard or soft-404 codes. One can also use the similar approach to find latest results and using them with the help of dead link detection algorithm against the external links present in an article. Following parameters are important while running such algorithm: page fetching timeout should be set around 10 seconds; in case no response is returned by the webserver then the page is considered dead. Maximum number of redirects should be set around 5; in case more than 5 redirects are observed then the page is considered dead to generate random URLs in the parent directory of the webpage, the method proposed in [12] could be adopted.

ISSN – 1999-4974

D. Wikipedia Article's length Correlation with External links

According to Wikipedia policy, articles with several external links may have incomplete information about the subject as they point the reader outside of the Wikipedia domain for additional information. On other hand the article with less external links may provide complete information about the subject as user need very less effort to read additional information from external links.

Articles that contain less information about the subject should be enriched through the content of external resources. To verify this assumption which is the part of so called policy of the Wikipedia, one needs to analyze the correlation between the external links and the articles' length. In order to verify the claim that, "usually large amount of external links are present in an articles that have large contents size" a study is presented in [14]. In [14] it is reported that on average there are 200 links in a lengthy article (i.e. having around 800K characters) and 3 to 8 external links for medium length articles (having around 10K characters). It should be noted that medium length articles are around 70.6% of the entire articles base in the English Wikipedia. Further in [14] it is claimed that,

"The peak in the quantity of external links is implicitly imposed by Wikipedia, which upon detection of long link lists invokes a warning to editors and suggests alternative ways of linking. But still, the fact that long articles contain more links than short ones connotes a contradiction between the instructions given in Wikipedia linking guidelines and the way in which these are adopted in practice."

In particular, Wikipedia editors are instructed to point to the external resources only if they think that they are not providing the complete detail about some subject. But, this is not observed by their results reported in [14] and it is found that articles with large length usually have more links than shorter length articles. Similar fact is also observed in the Wikipedia featured articles.

III. PROPOSED FRAMEWORK

To evaluate the external link credibility of an article one needs to evaluate all external links present in the article. The external link credibility of an article is denoted by E_{cred} which is based on all external links pointed by the article. Credibility is formulated by the summation of credibility of each individual external link that is denoted as C_{link} as in (1).

$$E_{cred} = \frac{\sum_{i=1}^{E_{num}} C_{link^i}}{R_{val}}$$

Here E_{num} and E_{cred} represent number of external links and external links credibility of an article respectively. The R_{val} is the ratio based factor, calculated by (2) in such a way that it is at least equal to the E_{num}. The (2) incorporates the link to length ratio.

$$R_{val} = \frac{(A_{max} + L_{fact})}{C_{max}}$$

A_{max} is maximum credibility that an article can gain based on all of its external links. Here L_{fact} is the ratio between the length of article and number of external links. Where $C_{max}$is the maximum value of the credibility of an external link. Evaluation of this ratio is enormously important because if we ignore this factor then results may fluctuate or become biased towards those articles that have short article length and high numbers of external links. To avoid such fluctuation log is used to calculate L_{fact}. Here L_{fact} is the article length factor which is calculated using (3).

$$L_{fact} = \frac{\log(A_{len})}{\log(E_{num})}$$

Here A_{len} is the length of the article. A_{max} is calculated using (4).

$$A_{max} = E_{num} \times C_{max}$$

In (4) C_{max} is the maximum credibility of an external link. Evaluation of $C_{link}$is quite tricky task as it is not evaluated based on a single factor. Basically credibility of an external link is multi-faceted attribute which needs an in-depth analysis of different factors. An in-depth study is performed and discussed in following section to evaluate these factors. The credibility of external link C_{link} is evaluated using the (5).

$$C_{link} = V_{link} \cdot E_{cp} \cdot (E_{sp} + T_{rank} + E_{date} + EP_{info} + VT_{link} + L_{cat})$$

In (5) V_{link}, E_{cp}, E_{sp}, T_{rank}, E_{date}, EP_{info}, VT_{link} and L_{cat} are verifiability, purpose, spatial significance, rank, recency, author information, text verification and category of the external link source respectively. Detailed explanation of the fact, that these factors computed and important for C_{link} formula is explained in following sections. But here, at least, it could be observed that V_{link} and E_{cp} are two scaling factors while E_{sp}, T_{rank}, E_{date}, EP_{info}, VT_{link} and L_{cat} are additive factors. Which mean that if either of V_{link} or E_{cp} is zero then the entire credibility of the external will become zero.

A. Dead links and Containment Effect

The most fundamental factors regarding credibility of an external link are two which are described in interrogative way as follows:

- Does the external link URL is valid (no dead or rotten link)? Does the referred text in an article exist or matches with the content provided by referenced external link?

A detailed architecture diagram of the proposed framework to calculate credibility of an external link is presented in Fig. 1. The proposed framework uses four external sources for this purpose namely, World Wide Web (WWW), Alexa data source using Alexa API, Google data source using Google API and MediaWiki data source using MediaWiki API. WWW is used to detect presence (dead or alive) of the referred link, referred text containment, recency of information, its domain and purpose while Alexa API helps in measuring link rank, owners information and its

ISSN – 1999-4974

spatial value. On the other hand MediaWiki and Google APIs help in evaluation of link type, country detection, and user recommendation and testimonials respectively.

To evaluate both factor namely whether "the referred link is dead or alive" and "the referred text in the article exist or matches with the content provided by referenced external link" one needs the value of verifiability denoted as V_{link}. Where V_{link} is formulated in (6).

$$V_{link} = D_{link} \times R_{cont}$$

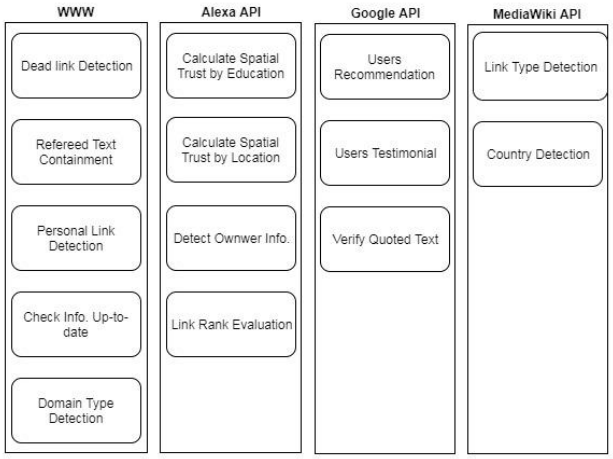

Fig. 1 Framework for Evaluating Credibility of External Links in Wikipedia

We need to evaluate the two factors namely D_{link} and R_{cont}. Here D_{link} is calculated through the dead link existence algorithm outcome which tell whether the link exist over the web or not. To evaluate this factor we implemented the algorithm presented in [12] to find the dead links. This algorithm also evaluates both soft and hard dead links. The containment of the referred text in provided reference external link is denoted as R_{cont}. To evaluate the containment we implemented the algorithm provided in [12]. This technique measures the syntactic text similarity which uses shingles [13] and evaluates the document similarity and containment of text in other document.

To find the referred text containment in an actual Wikipedia article is not a straightforward task as there is no single format to provide the references and people may follow from different provided methods to refer some text. Hence, some heuristics are used to formulate this task which is based on different types of external links as expressed in MediaWiki [17] markup format e.g. External links, References and Websites.

Wikipedia data is marked data that has tags in it so first we excluded all such tags to find pure text. Then we fetched complete text of the sentence followed by reference external link. Here we used another heuristic that if the text is too short then we fetched another consecutive sentence followed by a reference external link. Threshold for short sentence is taken up to 25 characters, so that if sentence is less than this threshold, then specific routine is triggered to fetch another sentence prior to the already examined one. Containment cont is the value between range [0, 1] and we used 0.5 as a fulfilling condition threshold to calculate R_{cont} using following formula.

$$R_{cont}\ (cont) = \begin{cases} 0, & \text{cont} < 0.5 \\ 1, & \text{cont} \geq 0.5 \end{cases}$$

V_{link} is a Boolean variable, if it is one then it means external links contains related text otherwise being zero it indicates external link is not related to the article and it nullifies all other verifiability of the external link. The reasons behind taking this factor as a Boolean variable are as follows:

- If the link does not exist on the web then there is no reason to calculate other factors to evaluate the credibility of an external link.
- If the referred text in the article does not exist on the source page of an external link then there is no reason to calculate other factors for credibility evaluation of the external link. This is managed using the text containment factor.

B. Effect of Education, Location and Web Traffic to external link: A Spatial Analysis for External links Web Traffic

The spatial factors may have huge impact on information credibility. External links that are regularly searched from graduate schools or colleges give us impression that they are usually referred in academia for learning or educational purposes, so the probability about information validity of such external links increase. Another important factor for spatial analysis of external links is that, if the external link belongs to same country from which article is assumed to be belong, then in such cases probability of valid information also increases because information is most likely to fulfill completeness, validity, freshness and may also include local or other geographical aspects. Such spatial effects of external link are calculated in form of external links spatial importance that is represented as E_{sp}. We formulated E_{sp} using (7).

$$E_{sp} = E_{edu} + E_{loc} + E_{area}$$

Here $E_{sp} \in [0, 3]$. E_{sp} depends on the spatial search of an external link based on E_{loc}, E_{edu} and E_{area} which have value either 0 or 1. Here E_{loc} represent external link's search by location e.g. work, home and school. E_{edu} represents whether external link is search by educational institutions e.g. college and graduate school and E_{area} represents whether the external link domain's country match with the country from which article's assumed to belong. The first two factors E_{edu} and E_{loc} are calculated on the basis of information acquired using statistics provided by the Alexa API [18]. The calculation of the E_{area} is not straightforward task and it is also not applicable on all articles. The reason behind limited applicability of $E_{area}$is because every article does not necessarily belong to a specific country e.g. the technology articles are usually not

 ISSN – 1999-4974

bound to a specific country rather they belong to the entire world. We use some heuristics for its evaluation; one heuristic which is used, is to fetch all the categories from which article belongs and then find the country from the category text. If we got more than one country from categories then we will use the country name with maximum occurrence in the categories of article.

C. External links Categorization by context

Category to which external link belongs is denoted by L_{cat}. It basically describes the category and type of website URL and domain that external link represents. External link category L_{cat} is evaluated through (8).

$$L_{cat} = T_{link} + NP_{link} + E_{dom}$$

Here type of external link is denoted as T_{link}. It has a fundamental importance as an external link from a book or a research repository has more chance to have valid information than any commercial website. Hence, if the type of external link is provided in metadata of the Wikipedia page as of a scholarly work then it is considered a well categorized external link which gives T_{link} a value 1 otherwise 0. The NP_{link} defines that whether an external link is someone's personal site or not. NP_{link} has value 0 if it is a personal website and has value 1 otherwise. In process of evaluating the information credibility of an external link it is an important factor. This factor encapsulate whether information on external link is biased towards someone personal affection or not. In case of a personal site there is a chance that information is biased toward some specific subject and it does not fulfill the neutral point of view policy imposed by the Wikipedia. The NP_{link} covers these factors. E_{dom} is the type of domain to which external link belongs e.g. edu, org etc. It is highly probable that external links which belong to an educational institute give more credible information than a commercial website. Hence if the domain of an external link is edu or org then the value of E_{dom} is 1 and otherwise 0. Hence, based on above $L_{cat} \in [0, 3]$.

D. External links Multi-Attribute Ranking

Another important factor regarding the evaluation of credibility of external link is the rank of external link in comparison to all other websites present on the WWW. To evaluate one of the attributes for ranking, we used heuristic based on popularity of some particular website among all others websites. Although it does not guarantee the information credibility of a website but it indirectly indicates the user/peer feedback in the form of its rank or popularity. This multi attribute rank of website is calculated using (9) and denoted by T_{rank}.

$$T_{rank} = GR_{link} + TR_{link} + UR_{link} + UT_{link}$$

The GR_{link} is the rank of external link with respect to all other web sites present on the WWW. Its value is calculated using its exact global rank given by the Alexa API which is transformed into range [0, 1] using (10).

$$GR_{link} = \frac{GR_{max} - GR_{act}}{GR_{max}}$$

We take the value of maximum global rank GR_{max} as 10,000, based on assumption that popular websites usually reside in this range of popularity. If a website does not reside in this range then its global rank has no effect on credibility evaluation of an external link pointing to it. Here it should be noted that if a website has global rank 1 then it is more popular than the website which has global rank 2. The GR_{act} is the actual global rank of a website based on its popularity with respect to all other websites retuned by the Alexa API. From (10) it could be observed that the website which has rank 1 on Alexa would has GR_{link}value almost equal to 1 which is 0.9999 and as the rank of a website gets closer to 10,000, than the GR_{link} gets smaller and eventually diminish to 0.

TR_{link} is trusted rank which tells that the external link is added in some user trusted search engine or repositories. User or peer recommendation is also taken in consideration regarding some specific external link. This factor is calculated as UR_{link}that is user recommendation outside the domain of Wikipedia for some particular external link. We calculated UR_{link} factor with the help of Google search API [19]. Another such similar factor is user testimonial to some particular external link in the form of blogs we see that either some external link is recommended in blogs or not. This factor is known as user testimonial UT_{link}.

E. External links Content Verification and Page Information

To evaluate the external link's website content is a complex task that needs an in-depth study of several techniques regarding the information retrieval and Natural Language Processing. Scope of this task is determined by just evaluating the verifiability of quotes used in the content of external link's website that is denoted as VT_{link}. To evaluate VT_{link} content of URL is fetched and its HTML tags are removed from it to get pure text and then quotes are found from the content. After that using Google API [19] search is performed for each quotes individually and it is cheeked that either these quotes are referred by some other websites or not. Quote existence threshold is set to 3 which mean similar quote should be referred at least three or more time on other websites too. Then there is a bright chance that it is reviewed or acknowledges by external world also.

An external link is believed to be more credible if the pointing source has information related to the author of the source. Hence the variable EP_{info} represents this factor. If the source of external link has name, and contact of the author available then EP_{info} will have value 1 otherwise it has value 0.

F. External links Classification

One of the most important factor is source verification is to know the facts about the purpose of external source website. If the purpose of some website is to do publicity or sale/purchase then this thing is obvious that it will mostly say

ISSN – 1999-4974

good about the subject e.g. its point of view would be biased about the subject to gain the market attraction. So to evaluate these aspects, an implementation of the Naive Bayes classifier is used to classify such external links properly. To achieve this goal, websites are categorized in following categories/classes with respective weights.

- Information [1.0]
- Sale and/or Purchase [0.5]
- Parody and/or Fun [0.3]
- Others [0.7]

The classifier is trained according to above four classes and then evaluation of external links is performed using this classifier to get predicted class. In this process first the text of webpage pointed by external link is taken as input of the classifier and then the respective value is assigned to the factor E_{cp} based on the predicted class. Here it should be noted that E_{cp} is a multiplicative factor in the construction of C_{link}, hence the over value of the credibility of an external link calculated in the (5) is scaled by the value of this factor. If the class of the webpage pointed by the external link is classified as 1 then the additive factor $(E_{sp} + T_{rank} + E_{date} + EP_{info} + VT_{link} + L_{cat})$ of the (5) is multiplied by 1 otherwise the additive factor is scaled down according to the respective value of the class.

IV. EVALUATION AND RESULTS

According to the proposed framework, external link should be a valid working URL (not a dead URL) and cited information in the article should be contained in the external link material. On meeting these two conditions credibility of the external link can be calculated based on other attributes of the external link. According to the proposed model, good credible external link of the article should have following characteristics:

- It should not be dead link.
- Its text is verifiable and well matched with reference information.
- It is up-to-date.
- It is regularly viewed from educational institutions and work places.
- If article belongs to a specific country then its domain should match with the country code.
- Type of external link is marked in Wikipedia.
- It should have high global rank and has users' recommendations and testimonial.
- There should be contact information about the external link founders.

Table 1 provides the details of all attributes with their respective symbols, characteristics, ranges and contexts used in the proposed model.

TABLE 1. EXTERNAL LINKS CREDIBILITY ATTRIBUTES

Symbol	Characteristic	Range and Context
E_{link}	Wikipedia article external link	External link
R_{cont}	Referred text containment in E_{link} content.	[0, 1], Relevance of the link
D_{link}	Is referenced E_{link} is dead	[0 or 1], link is live or not
V_{link}	Verifiability of E_{link}	[0 or 1], Is link alive and relevant
T_{link}	Type of E_{link}	[0 or 1], Link Type e.g. book, journal, blog, web etc.
NP_{link}	E_{link} is personal web or not	[0 or 1], Reduce personal bias
E_{dom}	E_{link} Domain type	[0 or 1], Type of Domain, commercial, non-commercial
VT_{link}	E_{link} Text verification	[0 or 1], verify that quoted text was not altered or not
GR_{link}	Global Rank of E_{link}	[0 to 1], Popularity of the link
TR_{link}	E_{link} Trust Ranking	[0 or 1], Is the page rated well in a local country directory
UR_{link}	E_{link} User Ranking	[0 or 1], Search by Google API
UT_{link}	E_{link} User testimonial	[0 or 1], Are some blogs link to it?
E_{cp}	E_{link} Constructing purpose	[0 to 1], Purpose of the page, to Inform [1], Sell [0.5], Fun [0.3], Other [0.7]
EP_{info}	E_{link} Page Author	[0 or 1], Who wrote the page? Email, Name etc.
E_{loc}	E_{link} Search by Location	[0 or 1], Searched by Home, School and Work
E_{edu}	E_{link} Search by Education	[0 or 1], Searched from academic institutes
E_{date}	E_{link} Temporal effect	[0 or 1], Is information up to date on an external link?

In order to evaluate the proposed framework, a Wikipedia page titled "Car Hydraulics" [20] is picked randomly and processed to calculate its external links credibility. This page has 788 words and 10 external links. Each link is processed and respective values as proposed in the framework are calculated. External links credibility of each external link of the page is reported in Table 2. In Table 2 bold-grayed rows are representing derived variables of the proposed framework while others are independent variables calculated through World Wide Web (WWW), Alexa, MediaWiki and Google APIs.

ISSN – 1999-4974

TABLE 2. EXTERNAL LINKS CREDIBILITY OF WIKIPEDIA PAGE ON "CAR HYDRAULICS"

Elink	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
Ecp	1.0	0.7	---	---	---	---	1.0	0.7	1.0	0.7
NPlink	1	0	---	---	---	---	0	1	1	0
Esp	**3.0**	**3.0**	---	---	---	---	**2.0**	**1.0**	**1.0**	**3.0**
Trank	**2.8**	**3.0**	---	---	---	---	**1.0**	**0.0**	**0.0**	**3.0**
Edate	1	0	---	---	---	---	1	0	0	0
EPinfo	1	1	---	---	---	---	1	1	1	1
VTlink	0	0	---	---	---	---	1	0	0	0
Lcat	**2**	**0**	---	---	---	---	**1**	**1**	**1**	**0**
Eedu	1	1	---	---	---	---	0	0	0	1
Eloc	1	1	---	---	---	---	1	0	0	1
Earea	1	1	---	---	---	---	1	1	1	1
Tlink	1	0	---	---	---	---	1	0	0	0
Edom	0	0	---	---	---	---	0	0	0	0
GRlink	**0.8**	**1.0**	---	---	---	---	**1.0**	**0.0**	**0.0**	**1.0**
TRlink	0	0	---	---	---	---	0	0	0	0
URlink	1	1	---	---	---	---	0	0	0	1
UTlink	1	1	---	---	---	---	0	0	0	1
GRact	2106	2	---	---	---	---	1	Max	Max	2
Vmax	**1**	**1**	**0**	**0**	**0**	**0**	**1**	**1**	**1**	**1**
Dlink	1	1	0	0	0	0	1	1	1	1
Rcont	**1**	**1**	---	---	---	---	**1**	**1**	**1**	**1**
cont	0.6	0.8	---	---	---	---	0.7	0.5	0.6	0.7
Clink	**9.8**	**4.9**	**0**	**0**	**0**	**0**	**7.0**	**2.1**	**3.0**	**4.9**

In Table 2 it could be observed that in total four external links (3, 4, 5, and 6) out of ten are dead links, hence their value for D_{link} is zero which yields V_{max} and eventually C_{link} for these external links as zero. Therefore other attributes related to these links are not calculated. One can read rows of the Table 2 using notations mentioned in Table 1, for example, from second row of the Table 2, external links 1, 7 and 9 are classified as for information purpose while 2, 8 and 10 are classified as others. Externals links 1, 8 and 9 are not personal webpages while 2, 7 and 10 are personal pages etc. External link 1 is of howstuffworks.com, 2 and 10 are of youtube.com and 7 is of google.com which have Alexa rankings 2106, 2 and 1 respectively. Based on the C_{link} values of ten external links of the page from Table 2 overall credibility of the external links of the page are calculated using equations (1), (2), (3) and (8). Here A_{len}, E_{num} and C_{max} are 788, 10 and 12 respectively. Hence,

$$A_{max} = E_{num} \times C_{max} = 10 \times 12 = 120$$

$$L_{fact} = \frac{log(A_{len})}{log(E_{num})} = \frac{log(788)}{log(10)} = 2.897$$

$$R_{val} = \frac{(A_{max} + L_{fact})}{C_{max}} = \frac{(120 + 2.897)}{12} = 12.55$$

$$E_{cred} = \frac{\sum_{i=0}^{n} C_{link^i}}{R_{val}} = \frac{31.69}{12.55} = 2.524$$

As maximum value of external links credibility of a page could be 12 hence the percentage credibility $E_{\%cred}$ of the Wikipedia page titled "Car Hydraulics" could be calculated as follows:

$$E_{\%cred} = \frac{E_{cred}}{C_{max}} \times 100 = \frac{2.524}{12} \times 100 = 21.03\%$$

External links credibility of the Wikipedia page "Car Hydraulics" is 21.03%, which is low and could be improved further.

V. DISCUSSION

Verification and credibility of external links is very important to gauge the correctness of provided information on a Wikipedia article. We have proposed a detailed framework in section III to evaluate the verifiability and credibility of external links. According to the proposed framework an external link is verifiable if it is a valid working URL (not a dead link) and cited information in article is contained by the external link material. If a link is verifiable then we can calculate further attributes of external link to measure its credibility e.g. we analyze the existing traffic to external link to evaluate the characteristic of external link that how much it is used at colleges, universities or work places etc. We also checked whether the external link is a personal web or not, so that we can maintain the check on neutral point of view policy and avoid the biasness. External link's URL country domain is also important to identify that article belongs to some country or not; similarly recommendations and testimonial about the external links credibility are also evaluated from outside of Wikipedia domain. To compute these and related factors proposed in the framework, four external sources are used which include World Wide Web (WWW), Alexa data source through Alexa API, Google data source through Google API and MediaWiki data source through MediaWiki API. Beside these data sources and APIs, several heuristics and algorithms are used to assess the verifiability and compute credibility of external links in a Wikipedia article.

VI. CONCLUSION AND FUTURE DIRECTIONS

In this paper, we have proposed a framework that evaluates the verifiability and credibility of all external links present in a Wikipedia article. This external link credibility is computed by accumulating credibility of all external links of an article based on factors that assist in external sources verifications. The framework has been evaluated on Wikipedia pages and external link credibility is computed. As the value of external links' credibility resides in the range of 0 to 12, so the percentage credibility of the article's external links could also be computed. This could help to

ISSN – 1999-4974

predict the detailed authenticity of sources pointing outside the article. In future this tool would be used as a module in computational trust framework for Wikipedia, where with other factors, it will help to measure computational trust that a reader could have on Wikipedia articles.

REFERENCES

[1] "Wikipedia:External links," Wikipedia. 18-Dec-2017.

[2] L. S. Buriol, C. Castillo, D. Donato, S. Leonardi, and S. Millozzi, "Temporal Analysis of the Wikigraph," in Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, Washington, DC, USA, 2006, pp. 45–51.

[3] J. E. Blumenstock, "Automatically Assessing the Quality of Wikipedia Articles," 2008.

[4] B. S. M. B. Twidale, "Assessing information quality of a community-based encyclopedia," in In Proceedings of the International Conference on Information Quality, 2005, pp. 442–454.

[5] W. Emigh and S. C. Herring, "Collaborative Authoring on the Web: A Genre Analysis of Online Encyclopedias," in Proceedings of the Proceedings of the 38th Annual Hawaii International Conference on System Sciences (HICSS'05) - Track 4 - Volume 04, Washington, DC, USA, 2005, p. 99.1–.

[6] J. Kamps and M. Koolen, "Is Wikipedia Link Structure Different?," in Proceedings of the Second ACM International Conference on Web Search and Data Mining, New York, NY, USA, 2009, pp. 232–241.

[7] J. Giles, "Internet encyclopaedias go head to head," Nature, vol. 438, no. 7070, pp. 900–901, Dec. 2005.

[8] N. Kirtsis, S. Stamou, P. Tzekou, and N. Zotos, "Information Uniqueness in Wikipedia Articles.," in WEBIST (2), 2010, pp. 137–143.

[9] "Wikipedia:Link rot," Wikipedia. 18-Dec-2017.

[10] "Wikipedia:Database download," Wikipedia. 18-Dec-2017.

[11] "API:Main page - MediaWiki." [Online]. Available: https://www.mediawiki.org/wiki/API:Main_page. [Accessed: 18-Dec-2017].

[12] Z. Bar-Yossef, A. Z. Broder, R. Kumar, and A. Tomkins, "Sic Transit Gloria Telae: Towards an Understanding of the Web's Decay," in Proceedings of the 13th International Conference on World Wide Web, New York, NY, USA, 2004, pp. 328–337.

[13] K. Monostori, R. Finkel, A. Zaslavsky, G. Hodász, and M. Pataki, "Comparison of overlap detection techniques," in Computational Science— ICCS 2002. Springer, 2002, pp. 51–60.

[14] P. Tzekou, S. Stamou, N. Kirtsis, and N. Zotos, "Quality Assessment of Wikipedia External Links," presented at the 7th International Conference on Web Information Systems and Technologies, 2011, pp. 248–254.

[15] A. Z. Broder, S. C. Glassman, M. S. Manasse, and G. Zweig, "Syntactic Clustering of the Web," in Selected Papers from the Sixth International Conference on World Wide Web, Essex, UK, 1997, pp. 1157–1166.

[16] "Help:Formatting - MediaWiki." [Online]. Available: https://www.mediawiki.org/wiki/Help:Formatting. [Accessed: 18-Dec-2017].

[17] "MediaWiki Format," http://www.mediawiki.org/wiki/Help:Formatting, 2013.

[18] "Alexa Web Services - Global Web Traffic Metrics," Amazon Web Services, Inc. [Online]. Available: //aws.amazon.com/alexa/. [Accessed: 15-Dec-2017].

[19] "Custom Search JSON/Atom API | Custom Search," Google Developers. [Online]. Available: https://developers.google.com/custom-search/json-api/v1/overview. [Accessed: 15-Dec-2017].

[20] "Car Hydraulics" [Online]. Available: https://en.wikipedia.org/wiki/Car_hydraulics. [Accessed: 15-Dec-2017].

ISSN – 1999-4974