University of Malta

Department of Intelligent Computer Systems

Master Thesis

Zhenghan Hua

Assessing the Credibility of Wikipedia Articles

Supervisor: Dr Alexiei Dingli

European Master of Science in Human Language Science & Technology (HLST)

2011

Saarland University
Department of Computational Linguistics and Phonetics

Master Thesis

Zhenghan Hua

Assessing the Credibility of Wikipedia Articles

Supervisor: Dr. Valia Kordoni
Prof. Hans Uszkoreit

European Master of Science in Language and Communication Technology (LCT)

2011

Abstract

The aim of this thesis is to contribute towards the research in the field of the evaluation of web articles' credibility. Wikipedia was selected as the source of such articles firstly as this is profusely used by students as well as the general public when searching for information. Secondly, it claims to be an encyclopedia, albeit a free one that everyone can edit. It therefore is representative of the genus of web information.

After a thorough research of past works in the field of credibility, web credibility and the history and nature of Wikipedia a system was devised in order to evaluate over 200 Wikipedia articles and designate a Final Credibility Score for each article. In the theoretical part of this thesis, previous research about web credibility is discussed, and some natural language processing (NLP) techniques are presented.

A system was designed in order to provide a Final Credibility Score. It was designed in a manner that reassembles select elements used in past research, namely Text Similarity, PageRank and Word Count, and, especially with reference to the computation of text similarity, develops them in an innovative manner. This is done in order to provide further information regarding the im/possibility of achieving a valid credibility rating using these three elements.

The system incorporates tools available to general public, namely Google and Java. This was done as an ideal system would provide easily achievable, prompt and automated results. The latter was one of the main rationales behind the creation of the system created during this work on the basis of web users requiring immediate results. The consultation of an expert would demand time and resources beyond the accessibility and desire of the average web user who selected the use of the internet over that of say a public library.

Declaration

I hereby declare that this diploma thesis is my own work and where it draws on the work of others it is properly cited in the text.

31st March 2011 Zhenghan Hua

Acknowledgements

I wish to express my greatest gratitude to my advisors Dr. Alexiei Dingli (Malta) and Dr. Valia Kordoni (Germany) for their guidance, support and advice. I would like to say a big thanks to Dr. Chris Staff and Mr. Mike Rosner for their assistance and infinite patience.

I would like to thank Bobbye Pernice for her dedication and support.

Many thanks to Theresa Calleja who taught me how to write a good thesis in English.

I also want to thank one of my friend Ling Cai who has helped me through many a crisis. I'd like to thank all my classmates in Malta for never failing to encourage and offer a helping hand.

Table Of Contents

Table of Figures

List of Tables

Introduction

Motivation

Information-accessibility to the general public has, over the past few decades reached unsurpassed heights. Electronic media and tools, especially the internet are one of the leading contributors to this phenomenon. This is positive development in that any individual may readily inform him/herself on a myriad of topics previous available to only a select few with access to reference texts and the likes. Nonetheless, this is also a double-edged sword. The onus is now on the individual to determine whether or not the text s/he is reading is reliable. Standards adopted by libraries of good standing are not equated on the web with the result that information can very possibly be misinformation or disinformation. How to evaluate the credibility of the web or web content is a challenging task, a lot of research has been done on this subject.

The evaluation of web credibility is multidimensional. From the web site point of view, the design of the web, information structure and focus, advertising, the information source, citations and readability can all be the indicators of the page's credibility. From the web content point of view, information accuracy, information clarity, the factualness, comprehensiveness or completeness of the text are all components of credibility. From the users' point of view, past experience with the site, age, education and purpose of use all influence the perception of the web credibility.

Aims & Objectives & Results

Significant work has been done on web credibility and various kinds of methods have been adopted. One method to evaluate web credibility is the "checklist method", for example, the method used by Fogg et al.[1] in 2001. In the checklist method, lists of items are provided and the users are guided to evaluate the web page they are interested in. Items in the checklist include whether the information is current and whether it is

accurate, as well as the author's qualifications. Other methods include the integration of different natural language processing techniques to identify web credibility as used in WISDOM[2] and other such "visualizing system"[3].

This thesis aims to contribute to the pool of knowledge in the web credibility evaluation with the hope that, in the future, web users may easily and assuredly identify the quality of information they are consuming. Wikipedia, due to its prolific use by information seekers was selected as the focus of this study.

The objectives are:

1. Creating an automated system for the evaluation of Wikipedia articles' credibility
2. Establishing the credibility of Wikipedia articles
3. Selecting and combining different criteria used in previous work for evaluating web credibility and assessing the result of such a fusion
4. Establishing whether readily accessible, consummate and 'easily' used computer/web tools provide sufficient resources for the creation of such a system.

Approach and Results

In order to achieve the above the content of web articles and some natural language processing techniques are used to analyze the credibility of Wikipedia articles. In particular, we sought to establish the validity of claims contained in an article whose credibility is to be determined. We assumed that if the same claims appeared in other sources that are reliable, then these would contribute to the article's credibility. Some experiments are conducted, their results computed and a Final Credibility Score assigned. The entire process, the program used and the results obtained are then examined and appraised. It turns out that due to the approach we took our results do not accurately determine the Final Credibility Score. We have analysed the approach, to understand why we obtained these results, and we make several recommendations to improve our system's accuracy.

Structure

This thesis can be considered to be a report on the progression of the work carried out to carry out the above. It includes five chapters. Chapter 1 is the introduction. Chapter 2 is the literature review, where an analytical synopsis of works consulted and drawn from in the theoretical research phase of this study is given. Chapter 3 outlines and justifies the methodology used in the practical stage of this study. Chapter 4 is the presentation and evaluation of results obtained. Chapter 5 is the analyses and critique of the results drawn and the method used to reach them. Based on this the strengths and weaknesses of the system are given and future work suggested.

Referencing

In this thesis The Chicago Style[4] of referencing was used. It was preferred as it offers sufficient flexibility for works referred to in the course of this study, many of which were available in electronic form and were highly interconnected. Examples of ways in which the Chicago style is used are as follows:

For single author books:

Chicago Style:

Pollan, Michael. *The Omnivore's Dilemma: A Natural History of Four Meals*. New York: Penguin, 2006.[5]

This thesis:

Salton, Gerard. *Introduction to Modern Information Retrieval*. New York: McGraw-Hill, 1983.

For articles in print journals as follows:

Weinstein, Joshua I. “The Market in Plato’s *Republic*.” *Classical Philology* 104 (2009): 439–58. [6]

This thesis:

Luhn, H. P. “The automatic creation of literature abstracts.” *IBM Journal of Research Development* 2(2) (1958):159–65.

For websites the following structure was used:

Chicago Style:

McDonald’s Corporation. “McDonald’s Happy Meal Toy Safety Facts.” Accessed July 19, 2008. http://www.mcdonalds.com/corp/about/factsheets.html. [7]

This thesis:

University of Waikato. “Weka 3: Data mining software in java.” Accessed February 28, 2011. http://www.cs.waikato.ac.nz/ml/weka/.

All references made in the text may be found in the References section of this thesis. All efforts were made to identify all the information necessary to complete the references in their entirety. However, in some cases, such as the specific location in which a conference was held, this could not be found and was therefore omitted from the reference. Nonetheless enough information regarding the conference would have been given to satisfy the requirements of an academic study as is this.

Chapter 2 Literature Review

Introduction

This chapter offers an overview of previous research on credibility as a general concept, web credibility and Wikipedia credibility. Next, text summarization and document similarity are focused on with a review of some natural language processing techniques used for key sentence extraction and the computation of similarities between texts.

Related works on Web Credibility

Credibility Issues

The invention of the internet has revolutionized human life and society. Dissemination and access to information by means of news agency sites, university OPAC systems, blogs, satellite mapping sites such as Google Earth, social networking, email as well as business advertising and exposure are but a few of the ways in which this milestone in human history has changed virtually every aspect of human interaction. The US Dept of Education places a lot of importance on the internet and technology as a means of education. The point made on numerous articles and posts[8] throughout the site is that technology and the internet provides a very successful educational tool, as can be seen in Fig. 1.

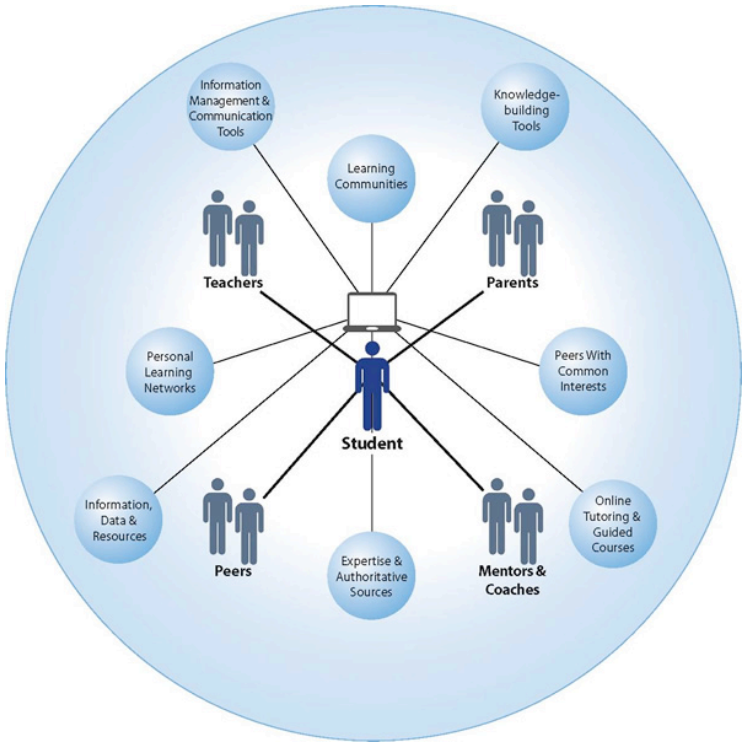

Fig. 1: *A Model of Learning, Powered by Technology*[9]

One feature that is not thoroughly dealt with is the exposure of students to sources of information that are not necessarily controlled by the teacher, i.e. a professional that disseminates quality i.e. credible information. Moreover, the 'computer' albeit internet sites in the figure above are not necessarily all fed by equally qualified and competent persons. In this case a statement made in the Federal Register of 4th June 2002, though not dealing directly with Web Credibility, highlights the necessity and scope of this research as well as others in the field, namely:

'Principal Federal statistical organizations are jointly announcing the opportunity to comment on their respective proposed guidelines for ensuring and maximizing the quality, objectivity, utility, and integrity of disseminated information.'[10]

In other words, if there is a lot of traffic towards the internet (most especially if educational institutions are initiating and/or encouraging such traffic) and a lot of information is being gleaned by this traffic (i.e. students), then, research in the field of information quality is of

the essence (students being a more vulnerable group that say professionals or adults with formed opinions and experiential information) to ensure that users can readily and easily identify what information is worthy of note and what is of less accuracy and integrity.

Web Credibility Projects

The question that therefore often presents itself is how trustworthy is information on the net. Access to the internet is world-wide[11]. Anyone with rudimentary knowledge of the system can put up a web page and can put any information s/he wants on it. As there is no international standard rating of every web site/web page it is often up to the user's discretion and ability to assess the value of content on a site s/he is viewing. This is not to say that no research has been carried out in the field. In the 90s The Stanford Web Credibility Project[12] started investigated possible ways of defining and asserting web credibility. Currently, one can view guidelines based on the results of their research in the field[13]. The guidelines largely reflect results from other research such as that by Fogg in 2001 and another in 2002. These studies shall be dealt with shortly within this chapter.

Other studies in web credibility include Credibility and Digital Media Project @ UCSB[14] which seeks to reach a better understanding of how users (professionals and otherwise) of a vast array of digital information resources perform searches for information and what are their preferred (albeit found to be more credible) sources of information. Consequently a thorough analysis of their understanding of credibility can be made with the aim of honing theories and methods of credibility assessment in present day digital media. Other bodies vesting considerable resources in the definition of such a notable and important concept include the Macarthur Foundation[15]. The latter considers the assessment of web credibility to be one of the important processes towards reaching their goal of 'building a more just, verdant, and peaceful world.'[16] Of the numerous grants the latter has dedicated to the role of digital media in education one finds $500,000 in support of an online immersive environment for young people to improve their media literacy skills and capacities at Indiana University (Bloomington, IN).[17] Web Credibility and Media Literacy are closely related, the former being an essential step once the latter has been achieved. As this goes beyond the scope of this study, more attention cannot be dedicated to this issue which is worthy of future research.

Credibility: Considerations & Definition

Nonetheless, despite these efforts, to date there is no clear, universally established and pertinent definition of credibility as a general concept, which is all too often dependent on the area/field in which and by which it is being defining. Nonetheless the overarching view is that credibility is:

> …credibility is the *believability* of a source or message, which is made up of two primary dimensions: trustworthiness and expertise. Some secondary dimensions include source dynamism (charisma) and physical attractiveness, for example. Moreover, the two primary dimensions (trustworthiness and expertise) have both objective and subjective components. That is, trustworthiness is a receiver judgment based primarily on subjective factors. Expertise can be similarly subjectively perceived but includes relatively objective characteristics of the source or message as well (e.g., source credentials or information quality)[18].

What applies to other fields cannot be transposed to the internet automatically. Hence the previous statement of the absence of one universally accepted and applicable definition of credibility. Burbules (2001) [19] gave the web's speed, complex features and link structure, as well as its lack of referencing and organizational conventions as reasons for the distinction of the internet from other sources of information. Concretely, while a reputable publishing house or journal will have policies and conventions governing the layout and referencing of publications/ contributions, there is no governing central body on the web to ensure that, say, references on a web page reflect all outlinks on the page. Therefore, when one reads a well-referenced and authoritative article on the web s/he cannot automatically infer that such a high standard of information can be assigned to other articles linked to that particular article. This issue shall be dealt with in greater depth later in this chapter in the section dealing with PageRank.

Despite the challenges we have just mentioned, research in the field of web credibility has been carried out. In a paper published in 1999, Tseng & Fogg[20], in one of their early works into credibility, started by giving a general outline of credibility. Today, twelve years later, with the skepticism, not to say cynicism, surrounding the internet and other media

the following may sound somewhat naïve and unattainable or utopian, still, it is worthy of note here as it is one of the foundations on which some of the leading researchers in the field of web credibility (as is Fogg) pivoted their research. Credibility is an expression of the unity of trustworthiness and expertise, Salwen & Stacks (eds.), 1996[21]. The latter they define as follows:

> Trustworthiness is defined as well-intentioned, truthful, and unbiased. The trustworthiness dimension of credibility captures the perceived goodness or morality of the source…Expertise is defined as knowledgeable, experienced, and competent. The expertise dimension of credibility captures the perceived knowledge and skill of the source. In other words, in evaluating credibility, a person assesses both trustworthiness and expertise to arrive at an overall credibility assessment.[22]

They proceed to define four kinds of credibility:

1. Presumed Credibility: Wherein the user/perceiver draws from preconceived notions/ suppositions regarding the authority of a source/person.
2. Reputed Credibility: A source becomes reliable based on reputation/reference from a third party
3. Surface Credibility: After a 'superficial' investigation of the source the user is content with assuming it is reliable.
4. Experienced Credibility: A source is deemed to be credible based on the assessor's first-hand experience with him/her/it over time.

The move on to logically infer that when assessing credibility a user draws from his/her expertise in the evaluation of credibility, his/her insight in the topic/area, his/her requirements for facts/data and finally the user's ability to gauge inaccuracies.

Based on the above, as well as from research to be discussed later in this chapter, the demarcation of web credibility is something of a minefield. The first question to be answered is 'credibility for whom'. In scientific research if the results of study cannot for instance be replicated or are not predictive then the study is deemed invalid. If a

Wikipedia article is claimed, following this study for instance, to be credible then it must necessarily ensure that the results of the study, i.e. the article is credible, should be identical to those of similar studies. However, as the definition of credibility largely depends, as illustrated by Tseng & Fogg above and by other researchers to be discussed shortly, on who is providing the evaluation i.e. is it an expert or not, is the person giving the assessment drawing from Surface Credibility or Experienced Credibility, what one study 'proves' to be credible another equally valid study may prove to be non-credible, even if both are using a randomly selected representative sample of the population of consumers. In a nutshell the main challenge in the definition of credibility is firstly the subjectivity of the definition/evaluation and secondly that research in the users of the internet is still in its early stages and, lest the situation is not complex enough, these are a very fluid 'constant', i.e. the demographics of users of the web is constantly and continuously changing making the selection of a representative sample for the scope of research a daunting task. These are but some of the questions underlying this as well as other studies in the field. In order to simply matter and narrow the scope of one's research, as shall be discussed in the Methodology Chapter, it is understandable to claim that an article is credible if experts in the field deem it such. However, if a source, such as Wikipedia, is designed for mass consumption/use, given that a less than a majority of the masses is in fact an expert, 'de factum' credibility is futile as consumers will not consider it such. Prior to delving further into the above quandaries a look at other research and studies in the area shall be given and the questions then explored further.

Web Credibility Studies

Despite possible reservations arising from the above considerations, initial research on web credibility was based on quantitative study. Later, with the development of natural language processing techniques some automatic methods on web information credibility were tried.

As discussed by Slater & Rouner in 1996[23] and later by Fragale & Heath in 2004[24] credibility assessments of sources and messages influence each other. If a user finds a source to be credible then s/he is highly likely to believe that the message is credible, this also applies to sources drawing from or referring to the original, positively evaluated

source. It therefore follows that if a web page is referenced by a reputable and trustworthy web page, such as an official government site, a famous university web site and/or a reputable company web site, then the web page is of importance and repute. Moreover, if the said web page is referenced by many other important and reputable web sites then it is even more likely that it is a quality web page. It should here be noted that when referring to 'quality', content rather than design, complexity, presentation or coding is taken into consideration.

In 2001 Fogg et al.[25] published a paper based on a large quantitative study. In the study, they investigated and outlined the factors which influenced the perceived credibility of websites. More than 1,400 people from the U.S. and Europe participated in the study and evaluated 51 different website elements. The elements could boost or hurt perceptions of web credibility. Finally they found five kinds of elements could boost the perceptions of web credibility which were "real-world feel," "ease of use," "expertise," "trustworthiness," and "tailoring." "Commercial implications" and "amateurism" were elements which could hurt the perceptions of the web credibility. This kind of study can help us to design the web more properly and then make it more credible. A year later, in 2003, Fogg et al.[26] drew up a report on how people evaluate a website's credibility. This report was also based on a large scale of study. In this study, 2,684 people were involved and they evaluated the credibility of two live websites on a similar topic. The comments people wrote about each site's credibility were collected and analyzed. Fogg et al. identified which features of a website were noticed when people assess a website's credibility. Finally they found the most frequently mentioned feature was "design look", followed by information structure and information focus. Fogg et al. also explored why the design look was so important when user evaluated the credibility of the website with Prominence-Interpretation Theory[27].

Stanford et al.[28] launched another study which was paralleled Fogg et al's 2002 study. In the study, experts from health and finance determined websites' credibility within their domains of expertise. The purpose was to compare the differences in criteria used by experts and consumers when they evaluated the same health and finance websites and to evaluate whether consumers failed in their assessments, and if so, how and where. Finally, they found the experts paid less attention to the visual appeal as a sign of web credibility and paid more attention to the quality of the website's information. The experts

were furthermore concerned about the source and references of the website as opposed to consumers who did not regard them as worth of too much note, if any.

News Credibility

From the research above we can again see that web credibility depends largely on who is assessing the said credibility. As Stanford et al.[29] above highlighted subjective factors are called into play when determining trustworthiness. While aesthetics play a key role to users, quality of referencing would override any visual considerations to professionals within the field. In an attempt to countermand the fluidity of subjectivities in 2002 Abdulla et al.[30] analyzed the components of credibility of news from newspapers, television, and online sites. The analysis was based on a national-wide telephone survey of 536 adults. Respondents were asked to evaluate the credibility of newspapers, television news, and online news. Twelve items (such as being up-to-date, fairness and honesty) were chosen and respondents were asked to rate the items on a five-point scale. Different dimensions were found when the respondents evaluated the credibility of different kinds of media. For example, balance, honesty and actuality were the main criteria for awarding of newspaper credibility.

Other efforts to evaluate news credibility automatically include that of Nagura et al.[31] who, in 2006, proposed a method to rate the credibility of news articles. The credibility of the news article was evaluated on three factors: commonality, numerical agreement, objectivity. The commonality was used to evaluate how similar the news article was to the other news articles which were collected from different publishers on the web. The numerical agreement was to check whether the numerical expressions in the original news article matched or not with the other collected articles. The objectivity was based on the speculative clue phrases and the indication of news sources. Every news article got a score based on these three factors. In the experiment, they got promising result: the average agreement between the proposed method and human assessments was 69.1%. Due to the success of Nagura et al.'s method, this study's methodology draws a lot from their research and system. This shall be further discussed in Chapter 3, Methodology.

In another study aimed at enabling online-news users to determine the credibility of an article, Kawai et al. (2008) [32] developed a 'visualizing news system'. The rationale

behind the study is that if a user can identify the trend, albeit bias, of an article then s/he can better 'translate' and identify the issues, extraneous to the information, that are being superimposed on it by the article's author/s. In Kawai et al.'s system a user enters a topic and news articles which include the keywords of the topic are retrieved from an already existing database. Sentiments are selected from the retrieved news articles and cross-referenced with a pre-compiled dictionary where the 'value' of the sentiment is rated on a four sentiment scale, each of which has its own 'range' as follows: Bright/Dark, Acceptance/Rejection, Relaxation/Strain, and Anger/Fear. A bar chart is then generated for every website where the user can clearly see how the topic in question ranks on the four sentiments in every website. S/he can then assess the trend of the article and decides on his/her own how credible the article is. Kawai et al.'s study differs from previously mentioned studies in that, rather than rating the credibility of a given piece, it presents elements visually to the 'consumer' who can then carry out his/her evaluation of the credibility of the topic in question. The value of such a method of assessment is that it somewhat mitigates the subjectivity factor i.e. users' different criteria for assigning credibility as shown in studies discussed above.

WISDOM

Another study that also puts the onus of attributing the 'final credibility score' on the user was carried out in 2009. An information credibility analysis system called WISDOM [33] was created wherein the information contents, information senders, and information appearances were used to estimate the credibility of information. When a query is sent to the system, WISDOM sends the query to a search engine. The top 1000 relevant web pages are retrieved for further analysis. Some natural language processing techniques (such as discourse analysis, key phrase extraction, syntactic analysis) are used to analyze the 1000 web pages. The user can view the analysis result by clicking some tabs. For example, the "Major/ Contradictory Expressions" tab shows a summary which includes the major phrases and contradictive phrases. The "Opinion" tab shows the analysis result of the evaluative expressions. The analysis result can help users to assess the credibility of the information related to the topic.

In Kawai et al.'s 2008 study mentioned and WISDOM, some natural language processing techniques (part-of-speech tagging, relation extraction etc.) were used to

assess information credibility. The methodology of this study draws from this concept as shall be discussed at length in Chapter 3, Methodology.

PageRank

This study's methodology is also built on another way of assessing the credibility of a web page that is unique to its genus, i.e. the world wide web, namely web link structure. This finds one of its key expressions in Google's PageRank. Nowadays Google has a toolbar that can be used to check the PageRank of every web site. PageRank can be utilized by a lay user to evaluate the significance of the article's web site.

In 1998, Page et al.[34] proposed the PageRank algorithm. PageRank is a popularity-based measurement of web page importance. The higher the PageRank is, the more credible the web page is.

PageRank can be expressed using the following formula[35]:

$$PR(A) = (1-d) + d(\frac{PR(T_1)}{C(T_1)} + ... + \frac{PR(T_n)}{C(T_n)})$$

where $PR(A)$ is the PageRank of web page *A*. When a random user surfs the web, at each step, there is a probability that s/he will continue to do so, d, the damping factor, is this probability and its value is between 0 and 1. The *d* is typically set to 0.85[36].

$T_i(i = 1, 2, ..., n)$ is the web page which refers to *A*, this too has its PageRank, which is taken into consideration. $C(T_i)(i = 1, 2, ..., n)$ is the number of outlinks of T_i. Therefore $PR(T_i)$ is divided by this value as PageRank of T_i contributes equally to these outlinks, which are independent of *A*, as it does to *A.*

As can be seen above, PageRank is highly dependent on numbers, i.e. the number of pages connected to a given site and their respective PageRank, which in itself is determined by the number of pages connected to them and their respective PageRanks. It is possible that PageRanking is increased by malicious websites by means of say, automated web site generating systems or by connecting reputable websites to their own. Nonetheless, when PageRank is incorporated with the query information, more accurate

searching results can be achieved.[37]

Credibility of Wikipedia

To Wiki or Not to Wiki

Yale College in its instructions on referencing to its students politely discourages the use of Wikipedia.

> Wikipedia certainly contains its share of incorrect information and uninformed opinion. And since it presents itself as an encyclopedia, Wikipedia can sometimes seem more trustworthy than the average website, even to writers who would be duly careful about private websites or topic websites. In this sense, it should be treated as a popular rather than scholarly source. [38]

Since there is such criticism and skepticism surrounding Wikipedia and since it is ranked as eighth in the top 500 sites used on the web[39] its popularity and credibility certainly deserve attention as it is plausible to suppose that traffic towards Wikipedia is typically in search of information. To date, the number of English written Wikipedia articles stands at 3,578,492[40], this number does not include the numerous articles written in other languages. Wikipedia provides a globally available and free resource. Painted against a backdrop of previously mentioned studies e.g. Fogg's 2001 criteria, it is no wonder that Wikipedia is a popular source of information worthy of note in its domain. However, due to the open approach (i.e. anyone can edit a Wikipedia article) whether the Wikipedia article is credible or not is a major concern of users, most especially academic ones as illustrated by Yale's reservations above.

Wikipedia Credibility Research

Of the numerable studies carried out on Wikipedia and Wikipedia articles the following works have been selected in order to provide some insight into the credibility of Wikipedia content. Different criteria have been used for these works.

In Lopes & Carrico (2008) [41] who claim that 'User interfaces play a critical role on the credibility of authoritative information sources on the Web.' This is due to the fact that

when using references as a means of evaluating the credibility of the source, as was the case with 'experts' in Stanford et al.'s study[42] a user must be able to gain access to these references easily. Understandably, a lack of access to references for evaluation is akin to poor credibility according to Lopes & Carrico 'This study has shown that there is a discrepancy on the accessibility of referenced Web pages, which can compromise the overall credibility of Wikipedia.'[43] In their study, Lopes & Carrico developed an automatic accessibility evaluation framework using checkpoints based on WACG[44] (Web Content Accessibility Guidelines) to check the accessibility features of the web HTML (HyperText Markup Language) structures. Every checkpoint contributed towards the final score which represents the accessibility level of the evaluated web page. The experiment was conducted on 365 web pages of which 100 were Wikipedia web pages and the others were Wikipedia reference pages. The experiment result proved that Wikipedia web pages have better accessibility than the reference pages leading to the authors to draw the conclusion cited above.

Quality and credibility are interdependent as stated by Rieh & Danielson based research carried out by Taylor (1986)[45] who coined six categories of user criteria for making choices: 'ease of use, noise reduction, *quality*, adaptability, time saving, and cost saving. He defined quality as "a user criterion which has to do with excellence or in some cases truthfulness in labelling" and identified five values included in quality: accuracy, comprehensiveness, currency, reliability, and validity.'[46]

Lim et al. (2006) [47] developed the basic model and peer review model. The models are based on the mutual reinforcement principle:

1. *Quality: An article has high quality if it is contributed by high authority authors.*
2. *Authority: A contributor has high authority if s/he contributes high quality articles.*

In the two models, the article quality is a function of the contributors' authority and the contributor's authority is a function of the articles' quality (the articles are the ones the contributor has edited). The quality value and the contributor authority can be computed iteratively like in HITS (Hyperlink-Induced Topic Search) [48] algorithm. The experiment results carried on the editing history of some Wikipedia articles shows that the efficiency of the two models to decide the article quality and contributor's authority. Lim et al.'s work provides us with insight as to how to use editing history to estimate the article quality.

Given that Wikipedia articles are dynamic, in that an existing article can be freely edited, a one-time ranking is not sufficient. An authoritative article can, if contributions that are not factual, accurate or authoritative, are made to it become fallacious or contain incorrect information. Consequently a continual assessment of the credibility of an article is required. Such an assessment ought to take into consideration not only, for instance, references and authorship, but also subsequent alterations to it.

In 2006, Zeng et al. [49] developed a revision history-based trust model which utilized rich revision information in Wikipedia to compute the article trust . This Zeng defined as follows "[the] r*evision history* of an article is a sequence of its versions ordered by their creation time"[50]. The model was represented with a dynamic Bayesian network[51]. The article trust value ranges from 0 to 1 wherein 0 means complete untrustworthiness and 1 means complete trustworthiness. Other values reflect how much percent content of the article is trustworthy. The test based on Wikipedia articles gave some promising results about the trust model which 'provided a methodology for comparing and evaluating future computational trust models.'[52]

Word Count Evaluation of Wikipedia Article Quality

In 2008, Blumenstock [53] proposed a simple method based on word count. Only the word count was considered in order to measure the Wikipedia article's quality. The articles were classified as two classes: featured and random. The task was to use article length (how many words it includes) as a feature and preformed a classification task on Wikipedia articles. 2000 words were chosen as the threshold to distinguish the featured and random articles. Even though the method was simple, high precision was achieved. The following figure (Fig. 2) is the word count threshold with the classification error rate. It can here be seen that 2000 is a reasonable threshold. This experiment has proved that article length is a valid means of assessing article quality. In view of this and the facility with which Wikipedia article length can be measured it has been chosen as a characteristic of this experiment as shall be explored further in the Methodology chapter.

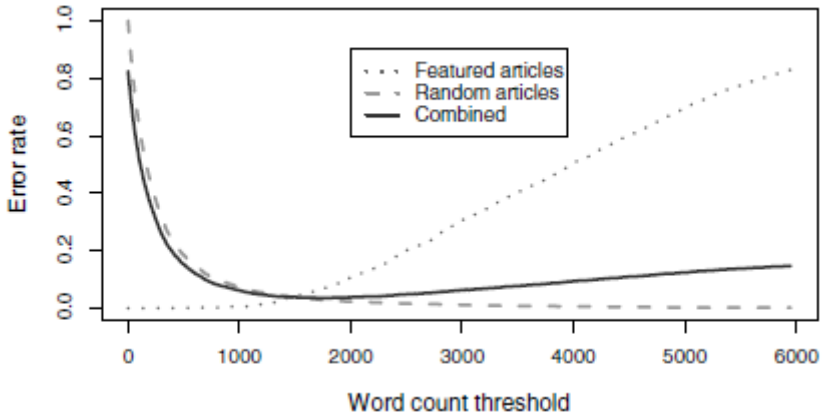

Fig. 2: Word Count Threshold[54]

Featured Articles

Another element of Blumenstock's[55] study which has been drawn from is this research is his selection of both Featured and non-featured Wikipedia articles. A featured article…

> … represents the best that Wikipedia has to offer. These are the articles, pictures, and other contributions that showcase the polished result of the collaborative efforts that drive Wikipedia. All featured content undergoes a thorough review process to ensure that it meets the highest standards and can serve as an example of our end goals. [56]

Non-featured articles, though possibly of valid content, have not been selected to 'showcase' Wikipedia and have therefore not undergone the 'thorough' review process. It can therefore be logically deduced that non-featured articles are less credible than featured articles. Wikipedia has an intermittent group between featured and non-featured articles called 'Good Articles'. For the purpose of this study these articles have not been selected as they would make the distinction between featured and non-featured articles less evident.

Other Research

Other research on Wikipedia includes Moturu and Liu[57] 2009 work which built a trust evaluation model and assigned a score for every Wikipedia article. The articles were

collected from the Wikipedia health articles. The trust evaluation model was based on features which included the proportion of paragraphs with citations, article size, factual accuracy, information on editing patterns, development history and user behavior, author credibility, revision count, proportion of reverted edits and proportion of reverted edits considered vandalism. The dispersion of the feature values from their mean was incorporated to assign a trust score for this feature ranging from 0 to 11. The scores from all the features were added together to provide the final trust score of the article. The larger the score is, the more trustworthy the article.

A different method for presenting final scores to users was developed in 2006 by McGuinness et al.[58] They presented how to present trust information to users in a visualizing way based on the trust values. Different article fragments have different trust values and the different trust levels were shown with different colours. The trust value is from 0 to 1 where 0 means unknown trustworthiness and 1 means complete trust. This visual representation about the trustworthiness of the article can help user to judge the credibility of the article.

One of the very latest works on the Wikipedia article quality is from Gabriel & Alex [59]. This is the first time that different models were used to compute Wikipedia article quality according to its category. Previous work, some of which has been discussed in this chapter, used the same model for all the articles. In 2010, Gabriel & Alex proposed quality models for the computation of Wikipedia article quality within different categories (six categories were assigned to Wikipedia articles: (1) stabilized articles, (2) controversial articles, (3) evolving articles, (4) list, (5) stub and (6) disambiguation page). The article quality score was computed based on its category and some specific features. Different features were chosen for different categories of Wikipedia articles, so, for instance, for stabilized articles features such as Log Length, Citation Density and Image Count Density were selected to compute the final quality score. Supervised machine learning algorithms provided by WEKA [60] were used to build the classifier and identify the category of the given Wikipedia article. In the evaluation part, only two categories: stabilized and controversial were used. 100 Wikipedia articles were chosen and 247 students (every student evaluated 8 randomly chosen articles) were asked to evaluate the information quality of the articles on a scale from 1 to 5. The average rating from the user for every article is represented as user opinion about the article. The test results

showed that the average user opinion and the quality score which was computed from the author proposed model had a positive relation.

Another recent work on Wikipedia article's trustworthiness analysis is from Lucassen & Schraagen (2010)[61]. 15 academic students participated in the experiment and every student evaluated 10 English written Wikipedia articles. Finally 120 articles were evaluated. Half of the articles were from a topic that was familiar to every user and the others were not from topic familiar to every user. The experiment result showed that the Wikipedia lay readers judge the trustworthiness mostly based on three categories of features namely text, references and pictures.

There have been various studies carried out in order to assess and evaluate credibility as has been illustrated earlier in this chapter. Some of which have been drawn from in this study as shall be discussed in Chapter 3, Methodology. A further technique used in this research is based on Sentence Extraction in order to identify key sentences in an article.

Sentence Extraction

Sentence extraction is a subtask of extraction-based text summarization. Usually some key sentences are extracted from a text and reorganized to generate a summary. Research on text summarization technique has been carried out for several decades providing and/or proposing various methods for sentence extraction as shall be demonstrated next.

In early days, some features such as word occurrences and topic relevance were considered in order to rank the sentence. In 1958, Luhn[62] used a method by means of which a score was assigned to every word (the words like "the", "a" etc. were excluded because their high frequencies) in the document. Then every sentence in the document was assigned a score according to a function of high frequency words which occurred in the sentence. The higher the sentence score is, the more important the sentence is.

On a complimentary note, in 1969, Edmundson[63] built an extracting system which was parameterized to control and vary the influence of the four components: key words, cue words, title and heading words, sentence location. As can be seen key phrase is here too an important factor to be considered for sentence extraction.

With the development of technology machine learning methods could be used for text summarization. Julian et al. [64] (1995) proposed a statistical classification method based on Bayes' rule. The features considered were sentence length cut-off feature, fixed-phrase feature, paragraph feature, thematic word feature and uppercase word feature. Given a training set of documents with hand-selected document extracts, the probability of a given sentence included in an extract was estimated. The sentences were ranked based on the probability and top scoring ones were retrieved as key sentences. In 2001, Conroy [65] proposed the hidden Markov model for text summarization providing a means for sentences extraction. In 2002, Nomoto and Matsumoto[66] proposed the method of probabilistic decision trees integrated with clustering framework. Another 2002 study was carried out by Osborne [67] who proposed the log-linear models to extract some sentences.

In 2004, Mihalcea[68] described a graph-based ranking algorithm in which every sentence in a text is taken as a node and the relation between two sentences is seen as an edge. On the sentence graph of the document, whether two nodes are connected or not is determined by the similarity of the two sentences. There are many ways to estimate the similarity between two sentences, but Mihalcea adopted the following similarity: suppose sentence S_i (where *i* is Natural number) is expressed with a set of N_i words which appear in the sentence: $S_i = W_1^i, W_2^i, ..., W_{N_i}^i$. Then the similarity between two sentences S_i and S_j is calculated as follows:

$$Similarity(S_i, S_j) = \frac{\left|W_k \middle| W_k \in S_i \,\&\, W_k \in S_j\right|}{\log(\left|S_i\right|) + \log(\left|S_j\right|)}$$

If the similarity is 0 between two nodes of sentences, then they are not connected. Otherwise, the two nodes of sentences are connected. Incorporated with these similarities, a weighted undirected graph is produced. Based on this weighted graph, the rank value of every sentence node can be recursively computed using the following formula[69]:

$$PR^W(V_i) = (1-d) + d * \sum_{V_j \in In(V_i)} w_{ji} \frac{PR^W(V_j)}{\sum_{V_k \in Out(V_j)} w_{kj}}$$

Where $V_i(i=1,2,\cdots,n)$ is the sentence node, d is damping factor like in the PageRank[70] algorithm mentioned earlier in this chapter. $In(V_i)$ is the set of sentence nodes which point to V_i. $Out(V_j)$ is the set of sentence node which are pointed by V_j. $w_{i,j}$ refers to the weight (in this case it means similarity between two sentences) between V_i and V_j. $w_{k,j}$ refers to the weight between V_k and V_j.

Finally the sentences with higher score can be selected. The experiment carried on some news articles selected from Document Understanding Evaluations 2002 [71] proved that the method was at least as good as the previously published results (DUC 2002[72]). The graph-based rank algorithm is suitable for single document summarization and an unsupervised method. No need to build the corpus manually like it is done in supervised learning method. Mihalcea's method and algorithm have been chosen to extract key sentences in our task as shall be further explained in Chapter 3, Methodology.

Text summarization research is still ongoing, as can be seen from the following mentioned works. Shareghi & Hassanabadi [73] (2008) proposed the Harmony Search-based sentence selection method for sentence selection. In 2009, Sankar & Sobha[74] identified coherent chunks in a text in order to rank them, they then found the sentences which had scores beyond a given threshold and used them to generate the summarization of the text. In 2010 Muratore et al. [75] used graph neural networks to perform the sentence summarization task.

Text Similarity

In the field of information retrieval, computing the similarity between query and document has been studied extensively. Many similarity methods have been proposed. The word based method, n-gram method and latent semantic analysis method[76] are all very renowned. The vector space model (which is used to index document) was first presented by Salton et al.[77] in 1975. Here, every document is expressed as a vector. For example, the text "tomorrow is a holiday" can be expressed as (tomorrow, is, a, holiday). If we just take the word occurrence into consideration, then the above vector can be expressed as (1,1,1,1). However, in large document set, the document vectors are used to compute similarity, so the term (word or phrase) weight needs to be reasonably

quantified since the words are not all equally important. One of the prominent term weight models is the tf-idf model presented by Salton[78] in 1983.

[79]Term Weight = $w_i = tf_i * \log(\frac{D}{df_i})$

where

- tf_i = term frequency (term counts) or number of times a term i occurs in a document.
- df_i = document frequency or number of documents containing term i
- D = number of documents in a database

The advantage of the tf-idf term weight model is that it incorporates the local and global information of term. We will use tf-idf term weight in our experiment. Other common similarity measures in information retrieval field were presented by Salton & McGill in 1981[80] and Meadows in 1992[81]. These measures are Dice, Jaccard, cosine, overlap, and asymmetric measures. We will use cosine similarity measure as shall be described in Chapter 3, Methodology.

Conclusion

Determining the credibility of information found in a web document is imminently desirable not only because of the increasing use of web information in education as Fig. 1 illustrates but also because, following the introduction of smart phones and other portable means of internet access, the web has become the most popular means of news consumption[82]. With reference to information consumption, though still not as popular as television and radio, a study carried out in 2009 by The Global Information Industry Center at the University of California proved that…

> ... computers have had major effects on some aspects of information consumption. In the past, information consumption was overwhelmingly passive, with telephone being the only interactive medium. Thanks to computers, a full third of words and more than half of bytes are now received interactively. Reading, which was in

> decline due to the growth of television, tripled from 1980 to 2008, because it is the overwhelmingly preferred way to receive words on the Internet.[83]

A credibility score, as this study presents or a bar chart with different elements of an article and/or web page for users to use when evaluating its credibility as done by Kawai and WISDOM, is therefore of the essence if people are to be aware of what it is that they are consuming, i.e. is the information reliable or is it, for instance, disinformation, used for political reasons as was the case during the Cold War[84]. Furthermore, users are also being given a 'new language'. As the Sapir-Wharf hypothesis postulates, behavior is heavily influences by the language being spoken[85]. The discourse and semantics of web documents are therefore also to be analysed and weighted for users to be able to identify more salient forms of credibility. Such analyses are necessary not only due to the human consumption of web information, which an individual is likely to pass on to another by word of mouth for instance, but also as automated systems and machines are increasingly availing themselves of web information both for the identification and for the analysis of web data. An example of the latter would be The Semantic Web[86].

This use and re-use of information is the motivation for research such as this and others mentioned earlier. As explained in the beginning of this chapter, web information is different from all other sources of information due to the web's speed, multiple-facets and link structure. Even more importantly is its lack of referencing and organizational conventions, Burbules (2001) [87]. This lack of referencing bears heavy burdens on the identification of sources of information. Nowadays if Stephen Hawking[88] were to make a claim on physics, this would be given high reliability scores due to Hawking having earning a notable credibility rating as a source following his innumerable prizes and awards and his twelve honorary degrees. Such a statement is consolidated by claims made by Salwen & Stacks (eds.), 1996[89] and Tseng & Fogg (1999) [90]. They postulate that if a source is credible then claims made by him/her/it are can be deemed credible. It can therefore be said to logically follow that if identical claims are made by an unreliable source, the latter can be bestowed credibility on the basis of it sharing the same information as reputable ones.

This theory forms the rationale behind the text similarity mode of assessment of credibility in this study. Wikipedia articles were summarized in order to identify key sentences and these were compared to reputable sources of information on the web. The identification of the reputability of such sources posed problems as shall be discussed in the Evaluation chapter to this research. As the goal of this study is to contribute towards the creation of an automated credibility-rating system, trust had to be bestowed on similar automated systems i.e. Google PageRank and Google returns. The discussion of such considerations and the postulation of what an ideal system would be, it is hoped, will make the latter's creation more immanent.

What is the use of an ideal system that automatically determines credibility? For one, vastly referred sources of information such as Wikipedia are subject to constant change. A credible article may be edited at any point in time and such editing render a previous credibility score unrepresentative of the latest version of the article. An automated system would enable users to, with the click of button, reassess the reliability of information found within an article.

Drawing from research discussed in this chapter, this study goes about selecting three criteria in the evaluation of credibility and puts them together in order to achieve a final credibility score. Source reliability is considered to be the most eminent method for evaluating credibility and was therefore given the centre stage in this study by means of the determination of a Text Similarity Score between a Wikipedia article and other sources of information, intelligence gleaned from Mihalcea 2004 study[91] and Lim et al. (2006) [92] was put to use here. This was weighted at 80% of the final credibility score, the remaining 20% were divided equally between PageRank, as per Page et al.'s 1998 work[93] and Word Count as per Blumenstock's 2008 research [94]. In the following chapter, how these three decisive factors in the awarding of a final credibility score were computed is given.

Chapter 3 Methodology and Implementation

Introduction

In an idyllic situation an 'oracle' would exist where one could refer articles to and be told in a short time how reliable, albeit credible, that article is. To date, despite its necessity, as illustrated in Chapter 2, no such infallible verdict-giving system exists. It should here be borne in mind that even if an article is 98% accurate and reliable the remaining 2% would render the entire article non-authoritative and/or lacking in credibility. Such a system would therefore necessarily have to be able to identify this 2% and give reasonable grounds for considering it inaccurate/unverified/unreliable.

The scope of this study is to contribute towards the formation of such an 'Oracle'. This study does not purport to define and create such a system. What this study does do, is refer to past research, as illustrated in Chapter 2, Literature Review, and, based on their methods and findings, select variables, systems and elements found therein in order to create a new, tentative, credibility-evaluating system and procedure. This was done so as to assess the procedure and results produced by such a 'new' selection and re-organization of elements used in previous studies. The evaluation of the strengths and weaknesses ensuing from this selection and re-coordination of variables and stages in the formulation of a system for the evaluation of web page credibility, aims to provide further groundwork for future studies in the field.

This chapter sets about explaining the rationale behind the selection of variables and elements from past research and how these were pieced together. However, prior to that, a reiteration of the necessity or research in this field is warranted.

In Chapter 2, Fig. 1, an example of how new technologies are becoming increasingly central to education was given. Still within the same chapter an example of educational institutions' reservations about credibility, specifically Wikipedia's credibility was provided with reference to Yale University[95]. However, credibility is not only an issue with reference to education. In 1938[96] Orson Wells delivered his 'infamous' broadcast[97] leading to mass

hysteria and mayhem. A CBS light entertainment programme was interrupted with the following:

> "Ladies and gentlemen, we interrupt our program of dance music to bring you a special bulletin from the Intercontinental Radio News. At twenty minutes before eight, Central Time, Professor Farrell of the Mount Jennings Observatory, Chicago, Illinois, reports observing several explosions of incandescent gas, occurring at regular intervals on the planet Mars. The spectroscope indicates the gas to be hydrogen and moving towards the Earth with enormous velocity..."[98]

Nowadays, reactions to such a radio broadcast might be of a less naïve and more skeptical nature. However, if a not so reputable news agency had to, on the morning of March 15th 2011 report that following the tragic events in Japan a nuclear cloud were spreading to the States, one can guess that some drastic measures may have been taken by some more gullible audiences. Y2K was to be the end of the world, though no 'reliable' sources stated that it would be, extreme reactions, including group suicides ensued[99]. Consequently, the importance of assessing credibility and providing, in a timely manner a reliable 'verdict' is essential.

Evaluating Credibility

The problem, as explored in Chapter 2, is "'Credible' for whom?". As studies referred to Chapter 2 illustrate, what is credible to the lay user is not so for the expert or professional one. In 1999 Tseng & Fogg[100] gave four different principles on which credibility to a source is awarded. An analysis of Wells broadcast above shows how all four criteria were met:

1. Presumed Credibility: The audience's preconceptions on CBS, Intercontinental Radio News and other features of the broadcast lead them to attribute reliability to the source.
2. Reputed Credibility: The information was substantiated by reference to Professor Farrell of the Mount Jennings Observatory, Chicago, Illinois. Whether the source was fictitious or not was irrelevant, due to Presumed Credibility audience would not have questioned whether such a person existed.

3. Surface Credibility: Well's authoritarian voice, the interruption of an entertainment programm for a news bulletin, the structure of the news bulletin all gave a superficial impression of credibility.
4. Experienced Credibility: CBS had not, prior to this event, 'deceived' it's listeners in such a way.

For the scope of this study the first three of the above four criteria shall be integrated in the evaluation system in an attempt to provide a holistic and/or more representative evaluation of credibility. The fourth has not been selected as it would have introduced variables hard to integrate in an automated computerized system, though it is to some, albeit limited, degree taken into account in the calculation of PageRank as shall be further explored shortly.

Another issue arising from the genus of this system, i.e. that it is automated and not human, is the identification of claims made in a sentence/article. Sentence extraction using the algorithm used in this study, as shall be defined later within this chapter, does not pose great difficulty. Evaluating claims made within or by the sentence on the other hand requires either natural language processing techniques as used in WISDOM and explained in Chapter 2 or an evaluation of who it is that is making the claims and, on the grounds of verified Reputed Credibility 'assuming' that the claims are reliable as shown by Slater & Rouner's 1996 study[101] and restated by Fragale & Heath's 2004 work[102]. This study uses features of both as shall be explored shortly.

In addition, another facet of credibility that poses considerations for such a research is its 'dynamic' nature, it changes over time. Guttenberg, Germany's Defence Minister had to resign due to his credibility suffering a severe blow following the discovery of his having plagiarized his PhD thesis:

> He was voted Germany's most popular politician.... But
> Karl-Theodor zu Guttenberg has resigned as defence minister
> after being engulfed by a plagiarism scandal…[103]

The same can be said of Mubarrak-related information. In just a few days his position went from President of Egypt to his resigning office and handing over military power[104].

Credible news articles wanting to retain such credibility would consequently have to be constantly updated in order to correctly reflect current affairs. On a more scientific note, science being the least volatile of areas (unlike politics say) here too, with Pluto 'demotion' from one of the nine planets in the solar system to planet to a dwarf planet. Our solar system subsequently now has eight planets. Any article on astronomy that does not reflect and include such updates therefore becomes unreliable.

Such changes in 'information' are among the considerations that have to be taken into account when creating an 'ideal' credibility rating system. Therefore, prior to exploring the methodology of the system proposed by and used throughout this research, a definition of such an ideal system shall be given. As shall be illustrated shortly, contributions made from this study can assist in the creation, in future work, of this ideal solution.

Design of an Ideal Solution

In idyllic conditions, one ought to start by selecting an article for evaluation. Claims made within the article are then selected. As explained earlier the selection of claims is hard as, while sentences are clearly defined as such (they all start with a capital letter and end with a full stop, exclamation mark or question mark), a claim is less obvious and can be interwoven in a sentence in a manner that makes its presence somewhat ambiguous. Still, ideally, through, for instance, semantic analysis, lexical analysis and even numerical analysis all claims within a text, be they of an ambiguous nature or not, can be identified.

Once these have been identified, their reliability needs to be assessed. In perfect conditions, one would have the time and resources to refer to an expert of undoubted credibility and his/her 'verdict' would resolve the issue of whether the claims are reliable or not. This however takes time and money among other things. An automated system providing a 'verdict' of equal reliability would therefore be preferable. This too however poses some difficulties. Such a system can be domain specific. It can, moreover, require automated reasoning and/or natural language processing techniques, these, though improving by the day, cannot claim infallibility, neither can they compare to the expertise, flexibility and insight as well as foresight of a human professional.

Nonetheless, let us imagine that all claims have been verified by a credible source,

leading to a verdict of 'reliable'. Ideally, a record of each reliable source is kept so as to enable every user who is questioning the reliability of a claim to view the source. Moreover, claims in an article that cannot be verified as credible should be marked as such so a user can easily distinguish between article content that is to be relied upon and content that is not. Finally, based on the ration of verified is to non-verified claims a final credibility score is given to the entire article. What such a ratio should be is matter worthy of future research as has been listed in the Future Research section of this dissertation.

Approach taken

In order to evaluate our system, a total of 300 articles were chosen from Wikipedia's 'Geography & Places' section from the contents page[105]. Of these, 177 were featured Wikipedia articles and 123 non-featured Wikipedia articles. Of these 300 articles, 83 were removed from the test set reducing the total number to 217, due to Google's blocking queries, this shall be elaborated on in Chapter 4, Evaluation. Of these remaining 217 articles, 116 were featured and 101 were non-featured articles (the distinction between which was explained in Chapter 2, Literature Review).

The three attributes considered for the assessment of a Wikipedia article's credibility are PageRank, Word Count and Text Similarity. The whole process was implemented using Java since it is commonly used and has a high degree of flexibility.

Obtaining a Similarity Score

Of the three attributes mentioned earlier, a lot of importance was given to the computation of the Text Similarity Score, the reason for this is explained in The Final Computation further on in this chapter. The flowchart for the computation of the similarity score is as follows in Fig. 3:

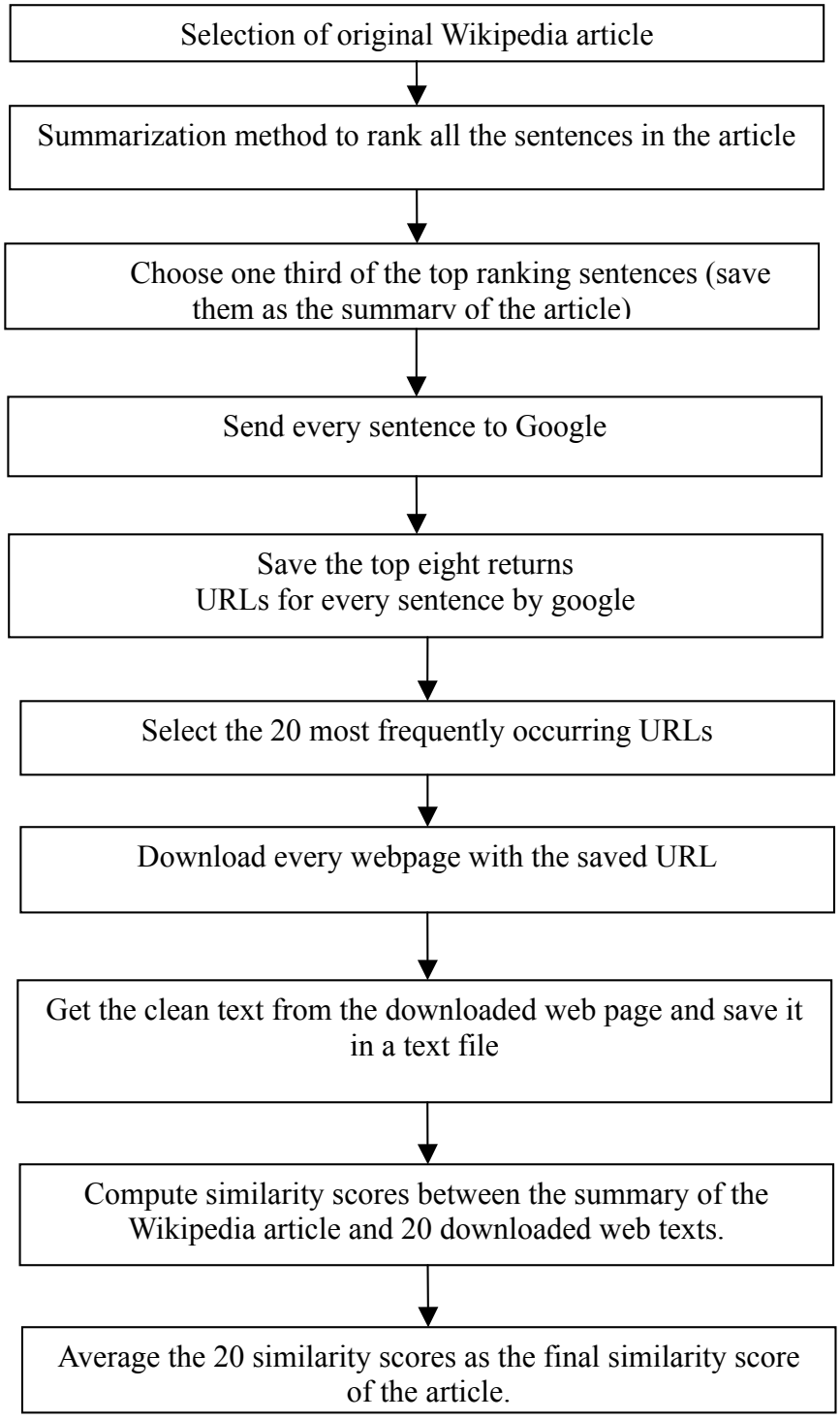

Fig. 3: Similarity Score Computation

Selection of Wikipedia Articles

Three hundred articles (reduced to 217 as explained earlier and further elaborated on in the Evaluation chapter) from Wikipedia were selected manually using cues based on

geographic details. Geography was chosen as a topic since it leaves less room for subjective claims and information. Topics of, for example, a more philosophical, economical, political, social, historical nature would have provided a less concrete basis for the verification process. Once the article was given by Wikpedia, the URL was used to find the HTML version of the page. An open source parser HtmlCleaner[106] is used to filter the Wikipedia article content so as to have the 'pure' text free from, for instance, HTML tags and java script. The content is now good enough to perform the following tasks.

Identification of Key Sentences

Once the content has been cleaned the article is split into sentences. This is necessary as key sentences are to be found. It is imperative that these sentences are the best possible 'representatives' of the main concept of the article. Moreover, it is 'assumed' that it is within the key sentences that claims to be verified are located.

Stanford MaxentTagger[107], a log liner Part of Speech Tagger, is used to split sentences. Though originally designed to identify parts of speech, it is good enough for the task at hand due to the fact that one of its features is sentence splitting.

Once the sentences have been split, key sentences are extracted. Since the article may be very long it is not practical to use the entire article as a query during a search for similar content on the web. A key sentence extraction method is used to identify key sentences. As reviewed in Chapter 2, many methods are used to locate key sentences in a text. Of these, Mihalcea's [108] graph ranking algorithm has been selected since, as explained in the Literature Review, it is well suited for single document summarization as well as an unsupervised method. It consequently makes an annotated corpus unnecessary, saving time and resources, both of which were precious in the course of this dissertation. Moreover, as the task is performed on Wikipedia pages dealing with geographical places the algorithm is suitable for the task as these are single-topic specific.

Once the sentences have been split, using Mihalcea's method they are given a ranking. The top one third of the sentences is chosen as the key sentences of the article. We

choose one third of the sentences because such a number of sentences are enough to represent the content of the article and within the fifty percent maximum given by Hovy (2005):

> … a summary is a text that is produced from one or more texts, that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). [109]

To Google or not to Google… to Google

Once the text has been summarized, the sentences within the summarization are compared to reliable sources to ascertain how similar they are and therefore determine whether or not the original article is in fact credible. This was chosen as a method for assessing credibility based on Slater & Rouner (1996)[110] and Fragale & Heath (2004)[111] studies as explained in Chapter 2.

It is here assumed that Google is a dumb oracle, it cannot tell us whether a claim is credible or not, but it can give us documents that may contain evidence that a claim is credible. In other words it is assumed that results returned by Google following a query contain only reliable, albeit relevant, sources in the top-ranking positions. Needless to say this is a very weak claim. If it were not so Yale's considerations mentioned in Chapter 2 and in fact this very study would be futile given that, during the pilot study stage of this research, out of 100 queries made an overwhelming majority yielded Wikipedia results in the Top 5 returns. When performing a Google search on anything from 'fish' to 'ashtray' to 'sociology' Wikipedia ranked first or second. When queries such as 'bank' were inserted then major banks worldwide were listed on the first page (12 or so results). So, for instance, Wikipedia ranked first on a query 'rottweiler' while 'www.dogbreedinfo.com/rottweiler.htm' [112] ranked second. As mentioned earlier, Wikipedia articles were selected based on geographical places. This choice was further consolidated by the fact that in the pilot study whenever a geographical query, such as 'Valletta', Paris', 'Berlin', 'Beijing' and 'Hong Kong' was entered in Google, Wikipedia always ranked first.

Whether information listed in Wikipedia was more credible than subsequent returns is hardly a matter for debate here. Having said that, in this pilot study carried out, the 'Surface Credibility' of articles featuring in the Top 5 returns to a query was more than adequate enough to provide part of the basis for evaluating the credibility of an article. Moreover, when the surface was scratched a little further no evidence to confirm that the top returns on Google were inadequate was found. This lent the researcher some 'Experiential Credibility', albeit in limited quantities with reference to the top returns.

Nonetheless, since the top relevance assessment criteria used by Google could not be identified, its role in the giving of the final credibility verdict exists solely within the parameters of a domain-specific search in the Text Similarity feature of this study, as shall be further explained shortly. Though this is not desirable in an idyllic system, it serves the purpose of this time and resource limited study whose final goal is to, as specified earlier, evaluate the feasibility and efficiency of combining elements of past studies in a new manner. Therefore, Google's criteria for returns in a query affect this study in so far as the computation of text similarity between a summary of Wikipedia article(one third top ranking sentences as summary) and the top 20 returns for queries using one third of the sentences within the original Wikipedia article.

Googling the Sentences

For the reasons listed above Google is used to get similar content for each key sentence, on the assumption that if each key sentence includes a claim that needs to be verified, and that Google returns only reliable sources in the top-ranking positions in the results set, then the documents that Google returns can be used to verify the claim. Every key sentence obtained from the article is submitted to Google without quotation marks, i.e. not as a phrase as one hundred percent replicas of the phrase are not necessary. Google Ajax search API[113] and JSON[114] library are used to automatically obtain the eight top ranking URLs. All the URLs are saved in a hash-table in which the key is the URL and the value is the recurrence of the URL. The same web page often returns more than once for key sentences within an article as the 'topic' would be the same.

Of these recurring URLs, the top 20 are chosen. This number was selected following a pilot study carried out on ten articles. It was found that after the top 20 recurring returns

for key sentence searches the number of recurrences per URL was insufficient to consider it as a similar enough article. It should here be noted that PDF URLs were excluded from the hash-table as they do provide encoded text. For every selected URL, the corresponding web page is downloaded and saved as an HTML file. The corresponding text is then extracted and cleaned i.e. stripped of HTML tags and java script etc. HtmlParser [115] is used to perform the task. This was preferred to HtmlCleaner[116], used for the cleaning of Wikipedia articles since, while the latter all have similar forms, different URLs found for the computation of similarity, do not.

Text Similarity

Once the texts of the recurring URLs have been cleaned the text similarity between the summary of the article and the texts is evaluated. The similarity is cosine value between two vectors. Stop words such as 'the', 'a' and forms of the verb 'to be' are removed because they do not carry much information about the content as per Luhn (1958)[117] study. The text vector's weighting is tf-idf measure as presented in the Literature Review. For the idf part, the number of the document set is 21(the summary of the article and the 20 similar web texts). Finally, for every Wikipedia article, the cosine similarity between the summary of the article (one third of the top ranked sentences in the article) and every text in the 20 similar txt files is computed. Then the 20 similarity scores are averaged as the final similarity score of the article.

PageRank

Once the text similarity has been scored, the article page's PageRank is identified. The PageRank value of every web page can be checked easily with a small java class.[118] Google provides a PageRank. Users would have to install the Google Toolbar on which a feature is present that presents users with the PageRank of the web page they are interested in. As this study requires the PageRank values of 217 web pages, manually logging of such PageRanks was not feasible. Consequently use was made of Google PageRank Java API[119] and Apache SoftWare[120] under licence 2.0.[121] in order to automatically get the PageRank value of a given web page. A text file was created manually in which all the URLs of the Wikipedia articles being evaluated were listed. Java program then read the file line by line (as each line is an individual URL) and obtain each URL's PageRank. This value is then divided by 10 since in its original state it is between 0 and 10. For the sake of the formula presented below in the Final Computation this value needs to be set between 0 and 1.

Word Count

As another criterion for the assessment of the Wikipedia articles' credibility is Word count, each of the 217 articles being evaluated had their words counted. This was done using a Java class. The Wikipedia articles were saved as a text file as a clean text. The file path was provided to the Java class which then counted the words. These then need to be given a weight between 0 and 1. As explained in the Literature Review, Blumenstock [122] gives 2,000 as a sufficient word count for the assigning of a high credibility score. Accordingly, if the word count for a given Wikipedia article exceeded 2,000 then this was given a weighting of 1, otherwise the weight was given as 0.9.These were then included in a hash-table for the final computation of the credibility of the articles.

The Final Computation

As explained in the Introduction to this study, the aim of this research is to contribute towards research carried out in the field of web credibility. Numerous studies, as shown in Chapter 2, have ventured on similar quests. To date no infallible, foolproof system exists. Lucassen & Schraagen (2010)[123] as discussed in Chapter 2, provided 3 criteria for the evaluation of Wikipedia articles. This study too uses three criteria. Since the scope of the system created here is to put together elements drawn from previous research, namely Text Similarity, PageRank and Word Count. The process for computing Text Similarity uses Salton's[124] 1983 concept for the representation of the document as a vector and the weighting of the terms used within the document. PageRank as used in this study, draws from Page et al.(1998)[125] and Word Count from Blumenstock's 2008 [126] work as well as Moturu and Liu 2009)[127]. These three criteria were given different weighting as shall be discussed shortly and computed to provide a final credibility score.

The Final Credibility Score (FCS)

Where TSS is Text Similarity Score, PR is the weighted PageRank and WC is the weighted Word Count, the final credibility score is computed as follows:

$$FCS = 80\% \text{ TSS} + 10\% \text{ PR} + 10\% \text{ WC}$$

The final score was given based on the summation of the three values (weighted to a value between 0 and 1 each) for similarity(TSS), PageRank(PR) and word count(WC). Text Similarity is given far more influence on the final score than the other two variables as this study computes the value itself and its veracity is considered to be sound, as it

was coined with reference to works by Nagura et al.[128] and Slater & Rouner (1996)[129]. Moreover it is possible to argue that PageRank, though of great validity, is vulnerable to malicious use as explained in Chapter 2. Word Count too can be 'doctored' by waffling on and on without making any substantial and/or original claims. Consequently, 80% of the value of the Final Credibility Score was attributed to Text Similarity.

The final presentation of the FCS is presented as a percentage. This is done on the basis that lay users are familiar with percentages and they can easily identify the credibility value of an article as being poor in the case of a FCS of 60% and high in the case of 85%.

Chapter 4 Evaluation

Evaluation metric

In natural language processing field, many kinds of measures are adopted in order to evaluate the performances of different tasks. For example, Bleu score[130] for machine translation, the Normalized Discounted Cumulative Gain (NDCG)[131] metric used in information retrieval system. Two commonly used metrics are precision and recall[132]. These were the desired evaluation criteria for our experiment. However, test results between featured Wikipedia Articles and non-featured Wikipedia Articles were not significantly different. This made the use of evaluation measures such as precision and recall not suitable for the evaluation of our results. Consequently, results were not assessed based on specific evaluation criteria. They were nonetheless analysed and yielded some interesting findings as shall be elaborated on shortly.

Time complexity

In order to judge the credibility of articles, the program used to compute the credibility score of every single Wikipedia article had to be run for a highly considerable amount of time, at times one article could take as long as twenty-eight minutes. This is due to the fact that the algorithm for computing sentence rank is laborious as many repetitions are required for the calculation of the rank of the individual sentences in the Wikipedia article. Their ranking had to be indentified in order to distinguish the top ranking sentences in the article. These were then used for sending queries to Google.

Another reason why the program took so long to run was that every web page in the search results had all of its content downloaded. This was additionally time-consuming as every line of HTML had to be read and written individually for further processing.

All other features of the program, such as, pre-processing and the computation of the similarity score were expedient.

The statistics of test data

As explained previously a total of 300 articles were selected using the process and rationale defined in the Methodology chapter. These were not divided equally between Wikipedia featured and non-featured articles since the purpose of this study is to evaluate credibility. As Wikipedia boasts of the quality of its featured articles, considering them to be prime examples of Wikipedia's work, these were given more prominence in this study.

The selected articles were analysed in terms of word count and sentence number as part of the key sentence identification procedure as explained in Methodology. Results for this analysis are as follows in Table 1:

	Number of Articles	Maximum Article Word Count	Minimum Article Word Count	Average Article Word Count	Maximum Article Sentences	Minimum Article Sentences	Average Article Sentences
Featured Wikipedia Articles	177	6401	419	2786	541	42	253
Non-featured Wikipedia Articles	123	7978	283	3217	662	27	280

Table 1: 300 Article Statistics

The statistics of the WC, sentence number etc. for the 217 articles are as follows in Table 2:

	Number of Articles	Maximum Article Word Count	Minimum Article Word Count	Average Article Word Count	Maximum Article Sentences	Minimum Article Sentences	Average Article Sentences
Featured Wikipedia Articles	116	6401	419	2929	541	42	386
Non-featured Wikipedia Articles	101	7978	283	3466	662	27	341

Table 2: Test Set Statistics

The formula used to compute the final credibility score, as explained in Chapter 3 is:

$$FCS = 80\% \text{ TSS} + 10\% \text{ PR} + 10\% \text{ WC}$$

The basic stages in the process as explained in Chapter 3 are illustrated in Fig. 4.

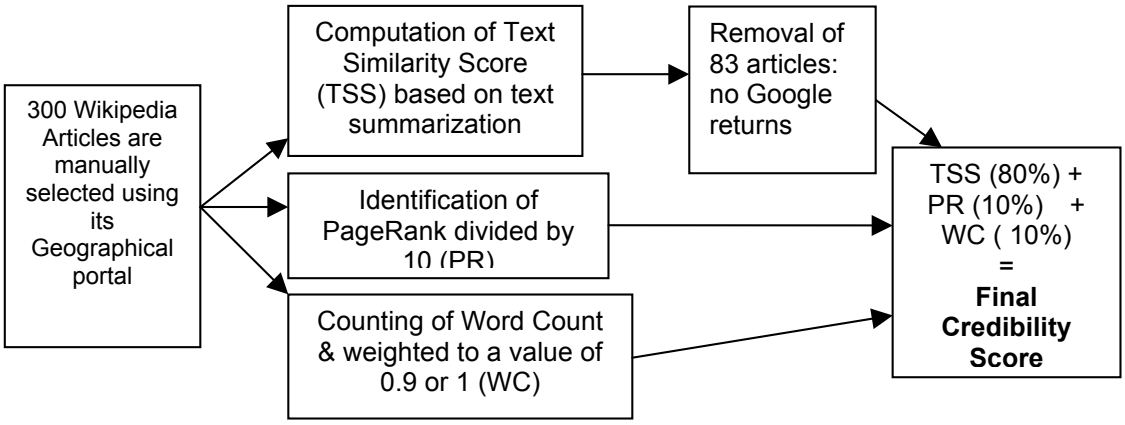

Fig. 4: FCS Computation

Test Results

In order to compute the TSS 300 articles were summarized and queries sent to Google of each sentence in the summarized text in order to identify other web pages with similar texts. Regrettably, Google has a system in place which blocks queries when these exceed a certain amount. Consequently when a few thousand queries were sent in a week Google would stop supplying returns for a query for a considerable number of days. The number of queries allowed and the number of days one would have to wait after being blocked varied. Needless to say, the process was greatly slowed by this.

Consequently, only 217 of the 300 articles selected could be given a TSS. The remaining 83 articles were therefore removed from the computation of the FCS since even though they had a WC, their inclusion would invalidate the FCS.

When computing the final credibility score, based on Wikipedia's claim that featured articles' 'content represents the best that Wikipedia has to offer'[133], such articles ought to be more credible and therefore results in the top ranking 116 articles should all be featured articles. In other words, if the system works correctly and Wikipedia's rationale[134] behind selecting articles to be 'featured' is sound, then the featured articles should have a high credibility score whereas the non-featured articles should have a low credibility score. As can be seen below this was not always the case. In the test results, 56 featured articles ranked among the top 116. So 48.3% of the 116 featured articles are ranked correctly. In the top 200 ranking articles, 53.5% is featured articles.

Test results	Number of Articles with FCS of 0.1+	Top ranked 116 Articles	Top ranked 200 Articles
Featured Wikipedia Articles	116	56	107
Non-featured Wikipedia Articles	101	60	93

Table 3: Test Ranking Results

Reasons for the discrepancy between Wikipedia's selecting articles as feature articles and the FCS resulting from this study can tentatively be said to be as follows:

1. Wikipedia does not continually evaluate their non-featured articles, in which case articles that will become featured once the evaluation process is carried out are currently not tagged as featured.
2. Wikipedia does not clearly label the processes an article is going through, it only gives the 'final label'. In other words, articles that are being evaluated or have not been evaluated (and therefore are not necessarily of inferior quality, but have just not been assessed) cannot be identified.

3. A featured article will continue to be such until “a consensus derived through discussion at the Featured article review page”[135] is reached. Such a consensus make take a while to be reached leading to articles retaining featured status while in fact they are not worthy of such a title. These too would have impacted the FCS of featured and non-featured articles.
4. No indication is given as to whether a featured article that has been edited retains its status or whether it is re-evaluated. Should a featured article be edited and as a result of changes made to it, there is a loss in its credibility, this too would impact the FCS by lowering the average FCS of featured articles.

Credibility Score Computational Pains

Below is a detailed breakdown of the FCS given by the system :

	Average FCS	Highest FCS	Lowest FCS
Featured Wikipedia Articles	0.1506756730781559	0.19489655666276914	0.11858422584359438
Non-featured Wikipedia Articles	0.15634169443101373	0.27891629948163177	0.1043408325555723

Table 4: Test FCS Results

Evaluation of Results

As is evident, the credibility score is quite low. This is due to the fact that the TSS of the Wikipedia article is low. The reasons for the TSS being low are as follows:

1) The similarity score is computed between the Wikipedia article and Google’s returning web pages’ web content. Some of the web pages returned by Google returning have less information, as a result of which the cosine similarity score is low. Accordingly, the average cosine similarity score between the summary of Wikipedia article and the twenty Google returned web pages is low.

2) The similarity measure is based on the cosine similarity of two vectors. Every element of the vector is the tf-idf weight of the term which occurs in the Wikipedia article and the Google returning web pages' content. This similarity measure does not capture sufficient semantic information. Using Lematizer and WordNet[136] would have yielded better results due to their ability to secure more semantic information.
3) Google returning web pages are based on the key sentences of the Wikipedia articles, the web page's content is less similar to the summary of Wikipedia article.

It should here be noted that the PageRank of every Wikipedia article's web page does not influence the finial credibility score significantly. Most of the web pages have a high PageRank (e.g. 7).

Reflections

Following the aforementioned issues and results the following reflections have to made:

1. Wikipedia articles appearing in Google's results set should be ignored. It logically follows that as results are being used to evaluate the credibility of Wikipedia then Wikipedia cannot be used to confirm and/or negate its own credibility. Regrettably an overwhelming number of the top returns for most of the searches made were in fact Wikipedia pages.
2. During the text summarization stage, key sentences in a Wikipedia article were identified so as to be submitted to Google for the election of top returns which are then to be used in the text similarity stage. At times the sentences were generic implying that though Google's result set contained evidence to endorse these claims, such an endorsement was not necessarily on the correct topic. For instance, many Web documents verify the claim "The visitors center is usually open annually from the Memorial Day weekend at end of May through Labor Day in September", but a result containing evidence of this is only relevant if it applies to the Visitors Center in Pithole, Pennsylvania. Such a flaw can be compensated for by expanding the query to include, for example, named entities in the title of the article or from the sentences close to the key sentence in the article.
3. Moreover, if Google's results can be trusted, only evidence from a single reliable

source that a claim is valid is needed. Merging results for each key sentence into a hash table and computing similarity to obtain the 20 most similar results was, if the above claim about Google's reliability is true, redundant. Not only that, it also resulted in 'losing' reliable sources that contained evidence for only one or two key sentences, implying that it did not feature in the top-20 documents returned. For example, it may be the case that 100 key sentences in the article whose credibility is to be determined are validated by 40 different documents retrieved by Google. However, after we identify the 20 documents out of those 40 that are most similar to the article whose credibility is to be determined, those 20 documents do not necessarily contain all 100 key sentences between them, so the Final Credibility Score may be lower than it should be.

Improving the System

1. Google cannot be as relied upon as was hoped. Search results were not satisfactory in view of the fact that, for instance, Google returned results from Wikipedia that, as explained in the Literature Review, are not necessarily credible. Google's site operator can be used to limit sources to authoritative articles in the geographic domain, such as http://www.nationalgeographic.com, http://geography.about.com.This could be repeated for other domains too.
2. Some web sites returned by Google contain information that is intended to be humorous e.g., Uncyclopedia (http://uncyclopedia.wikia.com/wiki/Main_Page), NewsThump (http://newsthump.com/). Obviously, it would have been inappropriate to use information contained in such web sites to judge the credibility of claims made in Wikipedia articles.
3. The process of determining if a trusted source verifies a claim contained in a Wikipedia article (or indeed any document whose credibility needs to be determined) can be refined as follows:
    a. Claims (or key sentences) in article are identified
    b. These are submitted as phrases to Google against a trusted collection
    c. If an exact match is found, then the claim can be verified
    d. If no exact match exists, the system would have to determine whether the claim has been reworded in the trusted document using approaches from

question-answering systems[137]

4. Some claims may require advanced reasoning or other approaches to solve. In other words a claim in a single sentence in one text may not necessarily find an identical counterpart in another text. Nonetheless the meaning of the claim may still be portrayed in a paragraph or more in another text, or across several documents (e.g., the source article contains a claim that the Prime Minister of Malta has 5 children. Two documents already determined to be credible that are being used to verify the claim contain only “Joe Borg has 5 children” in one document, and “Joe Borg is the Prime Minister of Malta” in the second document).
5. Claims may not clear. Anaphora, rhetorical features, as well as synonyms and homonyms can present difficulties to automated systems.
6. Some claims may be trusted because the author can be trusted, as Reputed Credibility, Tseng & Fogg (1999)[138]. This is so even though such claims do not already exist in other trusted sources, perhaps because this is the first time the claim is being made.

Conclusion

In this chapter, we have described our approach to evaluating our system. The evaluation results are not as accurate as we had expected, with no clear distinction being made between credible featured Wikipedia articles and non-credible not yet featured Wikipedia articles. Indeed, all Wikipedia articles in the test set are given a low Final Credibility Score.

We analysed the reasons for the low scores and made several recommendations to improve the mechanism for determining the credibility of individual claims (key sentences) in articles.

Conclusion

This work was successful in achieving its original aims and objectives to a reasonable degree. Its main goal of contributing towards the pool of knowledge available in the field of automated Web Credibility evaluation was reached based on the achievements listed further on, even when this was only in the identification of areas that require further work. The objectives as listed at the start of this work were (1) Creating an automated system for the evaluation of Wikipedia articles' credibility (2) Establishing the credibility of Wikipedia articles (3) Selecting and combining different criteria used in previous work for evaluating web credibility and assessing the result of such a fusion (4) Establishing whether readily accessible, consummate and 'easily' used computer/web tools provide sufficient resources for the creation of such a system.

The third objective was also reached in that the strengths and weakness of combining three elements taken from previous work and processed to produce a single Text Credibility Score (TCS). Results from formula computing Text Similarity, WordCount and PageRank were given a ratio value and fused in order to achieve a TCS. This TCS, though sufficient for the purposes of this study was hampered from reaching its full potential by limitations in the tools used. Google, for instance, limits the number of queries allowed in a given time frame. If an automated system is offered to the general public for the evaluation of web articles' credibility, Google would prevent it from performing after a number of queries are made, quashing its purpose. It has therefore been established that proficient readily and easily used web tools, such as Google, are, to date, inadequate for the operation of a pervasive system.

Determining credibility is a complex task. Our automated system, taking into account the recommendations for improving it, can establish that a claim in a source article is credible if the same claim appears (directly or indirectly) in one or more documents whose credibility has already been established (for instance, because they exist in an authoritative collection). However, if a claim is being made for the first time, then our system will be unable to determine its credibility. In this case, credibility would depend on other features, such as the credibility of the person who or entity that is making the claim;

the (future) validation of the claim made by trusted third-parties; the inclusion of a document containing the claim in an authoritative collection; and so on.

Further Research

Following the tribulations encountered throughout this study and the evaluation of its results, the achievements of this work are as follows:

1. Mihalcea's 2004 automated method for text summarization[139] was implemented yielding a reasonably dependable abridged version of an original text. This can be used for different scopes, such as the production of an overview for library purposes.
2. Claims are complex to identify and require further work and advanced systems to correlate. Homonyms and rhetorical features are especially taxing.
3. If a specific domain, e.g. Google, is going to be used to find authoritative articles to compute text similarity, the search must be restricted to certain key sites. These sites would have to be 'fed' to Google making the human element necessary.
4. Less is More: If a couple of sources, of indisputable credibility and different biases/agendas (e.g. SKY News and Al Jazeera rather than SKY News and BBC for reports on the Libyan Revolt), support a claim then more sources are not required. Limiting the number of sources used to verify a claim in an article would expedite the system and prevent being blocked by the search engine, as was the case with Google. The selection of these sources requires continual future research.
5. PageRank has been proved to be insufficiently dependable as a means of identifying authoritative websites and has to be supplemented using, at least another method for identifying credibility.
6. Further support has been given to reservations held by universities regarding the fallibility of Wikipedia as a source of credible information.
7. Different elements of a credibility assessment system should be given different weighting based on their 'standard error'. For instance, Word Count, while being a valid means of assessing credibility should be put in perspective in the sense that it can be manipulated and therefore should not carry as much weight in the final credibility score as, for instance, text similarity.

Based on the above the following Future Research has been identified as eminently desirable for the evaluation of Web Credibility and the providing of an accurate, sustainable and 'consumer-friendly' system:

1. Who says What… A study of Biases and Agendas in an Economic, Social and Political Information Universe

   If texts are to be used as references, be it as a primary source of information or in the computation of Text Similarity as a means of identifying credibility, biases and agendas need to be identified. Biases of say, a political party's station are easy to pinpoint. The agenda of a news network or even of an encyclopedia are less so. Such a study would provide a means, ideally automated, of content analysis in order to ascertain what the motivations behind a text are. It should here be noted that while National Geographic is highly authoritative in some fields its bias makes it an unreliable source of information in others. So, while it can be relied upon when giving the length of the Wall of China or some anthropological data it cannot be held as an expert on religious matters[140].

2. Credible for Whom?

   Religion is a key example of how a text may be entirely credible for one person and ludicrous to another, politics is another. On a similar note while Jung found Freud to be highly authoritative[141], Watson 'believed [it] to be highly subjective and unscientific.[142] Moreover, as shown in the literature review based on the expertise and education of a person evaluating the credibility of a text different criteria are used for assessing reliability, e.g. references in the case of a professional, layout and design in the case of a lay user. Such a study would seek to determining the different motivations one has when giving a personal 'credibility score' to a text, be it as a result of persuasions such as political affiliations or religious beliefs or, degrees of expertise or levels of education.

3. The Ultimate Buff Collection

Following ‘Who says What’ a collection of sources of information are given. Based on findings from ‘Who says What’ above, sources are listed and given as ‘Wholly Reliable’ or ‘Reliable in Matters of…’

4. On the Indisputable Nature of Expertise

   Who can claim to be an expert? Is there such a thing? Once an expert, always an expert? These are but some of the issues explored in this study along with: How incredible can you be? Who can you be incredible to?

5. The Missing Link: Of Rhetoric, Homonyms and others.

   If the human element is to be removed from Language Processing Techniques and Textual Analysis used in the evaluation of Text Similarity work to resolve considerations as follows is necessary. An automated system cannot easily identify whether a query for ‘sloth’ is done with reference to a mammal or a trait. On a similar note the use of the word ‘do’ differs based on the rhetoric being used. Neither can such a system currently easily identify and compensate for typos.

6. Credibility Awareness: You are what you Read/Hear

   This study would identify commercially viable ways of promoting credibility awareness and bias identification. It is now understood that ‘You are what you eat’, it is regrettably not equally understood that ‘You are what you read/hear’. Such a study would not only focus on academic text or news reports but even on schools books and popular culture.

7. Sentence vs Claim

   When reading texts in preparation for this study sentence and claim were often found to be interchangeable in works on text summarization and text similarity. Are they?

8. Text Summarisation: other forms & uses

Text Summarisation has reached advanced stages and can therefore be used for a myriad of purposes. Students can easily identify the desirability of their reading a text based on its overview. However, text summarization may also be advanced to produce simplified texts. A ‘simpler’ version of the text could then be used by, for instance, language students or readers who are not as proficient in the language of the original text as they would have to be in order to be able to read it with great loss of time and numerous frustrations. On a similar note, complex texts and articles could be summarized to accommodate young readers.

References

[1] B.J. Fogg et al. “What makes Web sites credible? a report on a large quantitative study.” Paper presented in the SIGCHI conference on Human factors in computing systems, 2001.

[2] Susumu Akamine et al. “WISDOM: A Web Information Credibility Analysis System.” Paper presented in Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, 2009.

[3] Yukiko Kawai et al. “Using a sentiment map for visualizing credibility of news sites on the web.” Paper presented in Proceeding of the 2nd ACM workshop on Information credibility on the web, 2008.

[4] The Chicago Manual of Style Online. **“**Chicago-Style Citation Quick Guide.**”** Accessed March 1, 2011. http://www.chicagomanualofstyle.org/tools_citationguide.html.

[5] Ibid.

[6] Ibid.

[7] Ibid.

[8] U.S. Department of Education. “Learning: Engage and Empower.” Accessed March 3, 2011. http://www.ed.gov/technology/netp-2010/learning-engage-and-empower.

[9] Ibid.

[10] Federal Register of 4th June 2002. Accessed March 3, 2011. http://www2.ed.gov/legislation/FedRegister/other/2002-2/060402d.pdf.

[11] “An Atlas Of Cyberspaces.” Accessed March 3, 2011. http://personalpages.manchester.ac.uk/staff/m.dodge/cybergeography//atlas/geographic.html.

[12] Stanford Web Credibility Research. “About the Web Credibility Project.” Accessed February 5, 2011. http://credibility.stanford.edu/.

[13] Stanford Web Credibility Research. “Stanford Guidelines for Web Credibility.” Accessed February 5, 2011. http://credibility.stanford.edu/guidelines/index.html.

[14] Credibility and Digital Media. “Credibility in the digital world. ” Accessed February 5, 2011. http://www.credibility.ucsb.edu/.

[15] Macarthur. “Overview.” Accessed February 28, 2011. http://www.macfound.org/site/c.lkLXJ8MQKrH/b.855245/k.588/About_the_Foundation.htm**.**

[16] Ibid.

[17] Macarthur. “MacArthur Awards $4 Million to Help Reduce the Threat of Nuclear Weapons.” Released February 17, 2011. Accessed March 3, 2011. http://www.macfound.org/site/c.lkLXJ8MQKrH/b.4196225/apps/s/content.asp?ct=9119185 .

[18] Andrew J. Flanagin and Miriam J. Metzger. “Digital media and youth: Unparalleled opportunity and unprecedented responsibility.” *Digital media, youth, and credibility* (2008): 5-28.

[19] Nicholas C. Burbules. “Paradoxes of the Web: The ethical dimensions of credibility.” *Library Trends* 49(3) (2001): 441–53.

[20] Shawn Tseng and B. J. Fogg. “Credibility and computing technology.” *Communications of the ACM* 42(5) (1999): 39-44.

[21] Self, C. C. “Credibility.” In *An integrated approach to communication theory and research,* edited by M B Salwen and Don W. Stacks, 421- 41. Mahwah, N.J.: Lawrence Erlbaum, 1996.

[22] Shawn Tseng and B. J. Fogg. “Credibility and computing technology.” *Communications of the ACM* 42(5) (1999): 39-44.

[23] Slater, M. D. and Rouner, D. “How message evaluation and source attributes may influence credibility assessment and belief change.” *Journalism and Mass Communication Quarterly 73* (1996): 974–91.

[24] Fragale, A. R. and Heath, C. “Evolving information credentials:The (mis)attribution of believable facts to credible sources.” *Personality and Social Psychology Bulletin 30* (2004): 225–36.

[25] B.J. Fogg et al. “What makes Web sites credible? a report on a large quantitative study.” Paper presented in the SIGCHI conference on Human factors in computing systems, 61-68, 2001.

[26] Fogg, B.J. et al. “How do users evaluate the credibility of Web sites? A study with over 2,500 participants.” Paper presented in the conference on Designing for user experiences, 1-15, 2003.

[27] Fogg, B.J. “Prominence-interpretation theory: Explaining how people assess credibility online.” Paper presented in Proceedings of CHI'03, Extended Abstracts on Human Factors in Computing Systems, 722-23, 2003.

[28] Stanford, J. et al. “Experts vs. online consumers: A comparative credibility study of health and finance Web sites.” Consumer WebWatch Research Report, October 29, 2002.

[29] Ibid.

[30] Abdulla, R.A. et al. “The credibility of newspapers, television news, and online news.” Paper presented at the Association for Education in Journalism and Mass Communication Annual Convention, Miami Beach, FL, August, 2002.

[31] Nagura R. et al. “A Method of Rating the Credibility of News Documents on the Web.” Paper presented in the 29th annual international ACM SIGIR conference on Research and development in information retrieval, 2006.

[32] Yukiko Kawai et al. “Using a sentiment map for visualizing credibility of news sites on the web.” Paper presented on the 2nd ACM workshop on Information credibility on the web, 2008.

[33] Susumu Akamine et al. “WISDOM: A Web Information Credibility Analysis System.” Paper presented in Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, 2009.

[34] L. Page et al. “The PageRank citation ranking: bringing order to the Web.” Technical report of Stanford Digital Library Technologies Project, 1998.

[35] Ibid.

[36] Ibid.

[37] Haveliwala, and Taher H. “Topic-Sensitive PageRank.” Paper presented in Eleventh International World Wide Web Conference, Honolulu, Hawaii, May 7-11, 2002.

[38] Yale College. “Wikipedia.” Accessed February 28, 2011. http://writing.yalecollege.yale.edu/wikipedia.

[39] Alexa. “Top Sites.” Asseed March 5, 2011. http://www.alexa.com/topsites .

[40] Wikipedia. “Main Page.” Accessed March 11, 2011. http://en.wikipedia.org/wiki/Main_Page.

[41] Rui Lopes and Luís Carriço. “ On the credibility of wikipedia: an accessibility perspective.” Paper presented on the 2nd ACM workshop on Information credibility on the web, 2008.

[42] Stanford, J. et al. “ Experts vs. online consumers: A comparative credibility study of health and finance Web sites.” Consumer WebWatch Research Report, October 29, 2002.

[43] Ibid.

[44] B. Caldwell et al. “Web Content Accessibility Guidelines 2.0.” Accessed March 1, 2011. http://www.w3.org/TR/WCAG20/ .

[45] Taylor, R. S. Value-added processes in information systems. Norwood, N.J.: Ablex, 1986.

[46] Rieh, S.Y. and Danielson, D.R. “Credibility: A multidisciplinary framework.” *Annual Review of Information Science and Technology* 41 (2007): 307-64.

[47] E.-P. Lim et al. “ Measuring qualities of articles contributed by online communities.” Paper presented in the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*,* December, 2006.

[48] Kleinberg, Jon. "Authoritative sources in a hyperlinked environment." *Journal of the ACM* 46 (5) (1999): 604–32.

[49] Honglei Zeng et al. “Computing trust from revision history.” Paper presented in Proc. of International Conference on Privacy, Security and Trust, 2006.

[50] Ibid.

[51] Zoubin Ghahramani. “Learning Dynamic Bayesian Networks .” *Lecture Notes In Computer Science* 1387 (1997): 168-97.

[52] Ibid.

[53] Joshua E. Blumenstock. “Size matters: word count as a measure of quality on Wikipedia.” Paper presented in the 17th international conference on World Wide Web, 2008.

[54] Ibid.

[55] Ibid.

[56] Wikipedia. “Portal:Featured content.” Accessed March 5, 2011. http://en.wikipedia.org/wiki/Portal:Featured_content.

[57] Sai T. Moturu and Huan Liu. “Evaluating the trustworthiness of Wikipedia articles through quality and credibility.” Paper presented in the 5th International Symposium on Wikis and Open Collaboration, 2009.

[58] D. McGuinness et al. “Investigation into trust for collaborative information repositories: A Wikipedia case study.” Paper presented on the Workshop on the Models of Trust for the Web (MTW'06), Edinburgh, Scotland, 2006.

[59] Gabriel De la Calzada and Alex Dekhtyar. “On Measuring the Quality of Wikipedia Articles.” Paper presented on the 4th workshop on Information credibility, 11-18, 2010.

[60] University of Waikato. “Weka 3: Data mining software in java.” Accessed February 28, 2011. http://www.cs.waikato.ac.nz/ml/weka/.

[61] Teun Lucassen and Jan Maarten Schraagen. “Trust in Wikipedia: how users trust information from an unknown source.” Paper preented on the 4th workshop on Information credibility, 2010.

[62] Luhn, H. P. “The automatic creation of literature abstracts.” *IBM Journal of Research Development* 2(2) (1958): 159–65.

[63] Edmundson, H. P. “New methods in automatic extracting.” *Journal of the Association for Computing Machinery* 16(2) (1969): 264–85.

[64] Kupiec, Julian, Jan O. Pedersen, and Francine Chen. “A trainable document summarizer.” Paper presented the 18th annual international ACM SIGIR conference on Research and development in information retrieval, 1995.

[65] Conroy, J. M. and O'leary, D. P. “Text summarization via hidden markov models.” Paper presented in Proceedings of SIGIR '01, 2001.

[66] Tadashi Nomoto and Yuji Matsumoto. “Supervised ranking in open-domain text summarization.” Paper presented at the 40th Annual Meeting on Association for Computational Linguistics, 2002.

[67] Miles Osborne. “Using maximum entropy for sentence extraction.” *Paper presented on* the ACL 2002 Workshop on Automatic Summarization*,* Philadelphia, Pennsylvania, July, 2002.

[68] Rada Mihalcea. “Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Paper presented in the 42st Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 170–73, 2004.

[69] Ibid.

[70] L. Page et al. “The PageRank citation ranking: bringing order to the Web.” Technical report, Stanford Digital Library Technologies Project, 1998.

[71] Document Understanding Conferences. “DUC 2002.” Accessed March 3, 2011. http://www-nlpir.nist.gov/projects/duc/data.html.

[72] Ibid.

[73] Ehsan Shareghi and Leila Sharif Hassanabadi. “Text summarization with harmony search algorithm-based sentence extraction.” Paper presented in the 5th international conference on Soft computing as transdisciplinary science and technology, 2008.

[74] K Sankar and L Sobha. “An approach to text summarization.” Paper presented in Proceedings of CLIAWS3, Third International Cross Lingual Information Access Workshop, Boulder, Colorado, June, 2009.

[75] Donatella Muratore et al. “Sentence extraction by graph neural networks.” Paper presented in the 20th international conference on Artificial neural networks, 2010.

[76] Deerwester, S. C. et al. “Indexing by latent semantic analysis.” *Journal of the American Society of Information Science*, 41(6) (1990): 391–407.

[77] G. Salton, A. Wong, and C. S. Yang. "A Vector Space Model for Automatic Indexing." *Communications of the ACM* 18(11) (1975): 613–20.

[78] Salton, Gerard. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[79] E. Garcia. “The Classic Vector Space Model-Description, Advantages and Limitations of the Classic Vector Space Model.” Accessed February 20, 2011. http://www.miislita.com/term-vector/term-vector-3.html.

[80] Salton, C, and McGill, M. Introduction to modern information retrieval. New York:

McGraw-Hill, 1981.

[81] Meadows, C.T. Text information retrieval systems. San Diego: Academic Press,1992.

[82] Mailonline. “News.” Accessed March 28, 2011. http://www.dailymail.co.uk/news/article-1366052/Internet-popular-news-platform-smartphones-tablet-use-grows.html.

[83] Roger E. Bohn and James E. Short. “How Much Information? 2009 Report on American Consumers.” Accessed March 2, 2011. http://hmi.ucsd.edu/pdf/HMI_2009_ConsumerReport_Dec9_2009.pdf.

[84] Garthoff, Raymond L. “Foreign Intelligence and the Historiography of the Cold War.” *Journal of Cold War Studies* 6(2) (2004): 21-56.

[85] Daniel Chandler. “The Sapir-Whorf Hypothesis.” Accessed March 18, 2011. http://www.aber.ac.uk/media/Documents/short/whorf.html.

[86] Semantic Web. “Semantic Web.” Accessed March 5, 2011. http://semanticweb.org/wiki/Semantic_Web.

[87] Burbules, N. C. “Paradoxes of the Web: The ethical dimensions of credibility.” *Library Trends* 49 (2001): 441–53.

[88] Prof. Stephen W.Hawking. Accessed March 20, 2011. http://www.hawking.org.uk/.

[89] Self, C. C. “Credibility.” In *An integrated approach to communication theory and research,* edited by M B Salwen and Don W. Stacks, 421- 41. Mahwah, N.J.: Lawrence Erlbaum, 1996.

[90] Shawn Tseng and B. J. Fogg. “Credibility and computing technology.” *Communications of the ACM* 42(5) (1999): 39-44.

[91] Rada Mihalcea. “Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Paper presented at the 42st Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004.

[92] E.-P. Lim et al. “ Measuring qualities of articles contributed by online communities.” Paper presented in the 2006 IEEE/WIC/ACM International Conference on Web Intelligence, December, 2006.

[93] L. Page et al. “The PageRank citation ranking: bringing order to the Web.” Technical report, Stanford Digital Library Technologies Project, 1998.

[94] Joshua E. Blumenstock. “Size matters: word count as a measure of quality on Wikipedia.” Paper presented in the 17th international conference on World Wide Web, 2008.

[95] Yale College. “Wikipedia.” Asseed February 28, 2011. http://writing.yalecollege.yale.edu/wikipedia.

[96] Gilbert Cruz.” A Brrief History of Orson Welles’ War of the Worlds.” Accessed March 22, 2011. http://www.time.com/time/arts/article/0,8599,1855120,00.html.

[97] Ourmedia. “The War of the Worlds.” Accessed March 22, 2011. http://www.archive.org/details/OrsonWellesMrBruns.

[98]Gilbert Cruz. “A BRIEF HISTORY OF Orson Welles' War of the Worlds.” Accessed March 10, 2011. http://www.time.com/time/arts/article/0,8599,1855120,00.html#ixzz1GezPXomy .

[99] Simon Dein. “Suicide and the apocalypse: a review.” *Psychiatry* 3(8) (2004): 21-23.

[100] Shawn Tseng and B. J. Fogg. “Credibility and computing technology.” *Communications of the ACM* 42(5) (1999): 39-44.

[101] Slater, M. D., and Rouner, D. “How message evaluation and source attributes may influence credibility assessment and belief change.” *Journalism and Mass Communication Quarterly 73* (1996): 974–91.

[102] Fragale, A. R. and Heath, C. “ Evolving information credentials:The (mis)attribution of believable facts to credible sources.” *Personality and Social Psychology Bulletin 30* (2004): 225–36.

[103] The Guardian. “German defence minister resigns in PhD plagiarism row.” Accessed March 20, 2011. http://www.guardian.co.uk/world/2011/mar/01/german-defence-minister-resigns-plagiarism.

[104] Aljazeera. “The rise and fall of Hosni Mubarak.” Accessed March 20, 2011. http://english.aljazeera.net/video/middleeast/2011/02/20112112120218476 46.html.

[105] Wikipedia. “Portal:Contents.” Accessed March 15, 2011. http://en.wikipedia.org/wiki/Portal:Contents.

[106]Htmlcleaner. Accessed March 15, 2011. http://htmlcleaner.sourceforge.net/.

[107] The Stanford Natural Language Processing Group. “Stanford Log-linear Part-Of-Speech Tagger.” Accessed March 15, 2011. http://nlp.stanford.edu/software/tagger.shtml .

[108] Rada Mihalcea. “Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Paper presented at the Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004.

[109] Hovy, E.H. “Automated Text Summarization.” In *The Oxford Handbook of Computational Linguistics*, edited by R. Mitkov, 583–98. Oxford: Oxford University Press, 2005.

[110] Slater, M. D., and Rouner, D. “How message evaluation and source attributes may influence credibility assessment and belief change.” *Journalism and Mass Communication Quarterly 73* (1996): 974–91.

[111] Fragale, A. R. and Heath, C. “Evolving information credentials:The (mis)attribution of believable facts to credible sources.” *Personality and Social Psychology Bulletin 30* (2004): 225–36.

[112] Dog Breed Info Center. “Rottweiler.”Accessed March 15, 2011. http://www.dogbreedinfo.com/rottweiler.htm.

[113] Google. “Google Code.” Accessed March 1, 2011. http://code.google.com/more/.

[114] Json. “JSON in Java.” Accessed January 3, 2011. http://www.json.org/java/.

[115] Htmlparser. Accessed March 2, 2011. http://htmlparser.sourceforge.net/.

[116] Htmlcleaner. Accessed March 15, 2011. http://htmlcleaner.sourceforge.net/.

[117] Luhn, H. P. “The automatic creation of literature abstracts.” *IBM Journal of Research Development* 2(2) (1958): 159–65.

[118] Wangcheng blog. Accessed March 20, 2011. http://wangcheng.javaeye.com/blog/411498.

[119] Temesoft. “Google PageRank Java API.” Accessed March 20, 2011. http://www.temesoft.com/google-pagerank-api.jsp.

[120] The Apache Software Foundation. “Apache download mirrors.” Accessed March 20, 2011. http://www.apache.org/dyn/closer.cgi.

[121] The Apache Software Foundation. “Apache Licence, Version 2.0.” Accessed March 20, 2011. http://www.apache.org/licenses/LICENSE-2.0 .

[122] Joshua E. Blumenstock. “Size matters: word count as a measure of quality on Wikipedia.” Paper presented in the 17th international conference on World Wide Web, 2008.

[123] Teun Lucassen and Jan Maarten Schraagen. “Trust in Wikipedia: how users trust information from an unknown source.” Paper presented on the 4th workshop on Information credibility, 2010.

[124] Salton, Gerard. Introduction to Modern Information Retrieval. New York: McGraw-Hill, 1983.

[125] L. Page et al. “The PageRank citation ranking: bringing order to the Web.” Technical report of Stanford Digital Library Technologies Project, 1998.

[126] Joshua E. Blumenstock. “Size matters: word count as a measure of quality on Wikipedia.” Paper presented in the 17th international conference on World Wide Web, 2008.

[127] Sai T. Moturu and Huan Liu. “Evaluating the trustworthiness of Wikipedia articles through quality and credibility.**”** Paper presented at the 5th International Symposium on Wikis and Open Collaboration, 2009.

[128] Nagura R. et al. “A Method of Rating the Credibility of News Documents on the Web.” Paper presented in the 29th annual international ACM SIGIR conference on Research and development in information retrieval*,* 2006.

[129] Slater, M. D., and Rouner, D. “How message evaluation and source attributes may influence credibility assessment and belief change.” *Journalism and Mass Communication Quarterly 73* (1996): 974–91.

[130] Papineni, K. et al. “BLEU: a method for automatic evaluation of machine translation.” Paper presented at the 40th Annual meeting of the Association for Computational Linguistics*,* 2002.

[131] K. Jarvelin and J. Kekalainen. “Cumulated gain-based evaluation of IR techniques.” ACM Transactions on Information Systems 20(4) (2002): 422-46.

[132] Christopher D. Manning and Hinrich Schütze. Foundations of Statistical Natural Language Processing. MIT Press,1999.

[133] Wikipedia. “Portal:Fearured content.” Accessed March 1, 2011. http://en.wikipedia.org/wiki/Portal:Featured_content.

[134] Wikipedia. “Wikipedia:Featured article criteria.” Accessed March 2, 2011. http://en.wikipedia.org/wiki/Wikipedia:Featured_article_criteria.

[135] Wikipedia. “Wikipedia:Former featured articles.” Accessd March 4, 2011. http://en.wikipedia.org/wiki/Wikipedia:Former_featured_articles.

[136] Princeton University. “Wordnet: A lexical database for English.” Accessed March 21, 2011. http://wordnet.princeton.edu/.

[137] Susan Dumais, Michele Banko, Eric Brill, Jimmy Lin and Andrew Ng. “Web Question Answering: Is More Always Better?” Paper presented in Proceedings of SIGIR\02, 2002.

[138] Shawn Tseng and B. J. Fogg. “Credibility and computing technology.” *Communications of the ACM* 42(5) (1999): 39-44.

[139] Rada Mihalcea. “Graph-based ranking algorithms for sentence extraction, applied to text summarization.” Paper preented at the 42st Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain, 2004.

[140] Faith Builders. “The National Geographic’s Judas Fraud.” Accessd March 4, 2011. “http://www.idscience.org/2008/05/17/the-national-geographics-judas-fraud/.

[141] Douglas A. Davis. “Oedipus Redivivus: Freud, Jung and Psychoanalysis.” Accessed March 20, 2011. http://www.haverford.edu/psych/ddavis/jungfreu.html.

[142] Education Encyclopedia. “John B. Watson (1878–1958) - Popularizing Behaviorism, The Little Albert Study, The ‘Dozen Healthy Infants’, Life after the University .” Accessed March 10, 2011. http://education.stateuniversity.com/pages/2543/Watson-John-B-1878-1958.html.