The current issue and full text archive of this journal is available at
www.emeraldinsight.com/1744-0084.htm

An experimental system for measuring the credibility of news content in Twitter

Hend S. Al-Khalifa
Information Technology Department,
King Saud University, Riyadh, Saudi Arabia, and
Rasha M. Al-Eidan
Computer Science Department, King Saud University, Riyadh, Saudi Arabia

Abstract

Purpose – Owing to the large amount of information available on Twitter (a micro-blogging service) that is not necessarily true or believable, credibility of news published in such an electronic channel has become an important area for investigation in the field of web credibility. This paper aims to address this issue.

Design/methodology/approach – A system was developed to measure the credibility of news content published in Twitter. The system uses two approaches to assign credibility levels (low, high and average) to each tweet. The first approach is based on the similarity between Twitter posts (tweets) and authentic (i.e. verified) news sources. The second approach is based on the similarity with verified news sources in addition to a set of proposed features.

Findings – The evaluations of the two approaches showed that assigning credibility levels to Twitter tweets for the first approach has a higher precision and recall. Additional experiments showed that the linking feature has its impact on the second approach results.

Research limitations/implications – The proposed system is experimental; thus further experiments are needed to prove these findings.

Originality/value – This paper contributes to the research on web credibility. It is believed to be the first which provides a proposed system to evaluate the credibility of Twitter news content automatically.

Keywords Blogs, Information media, Communication, Trust, Twitter, Credibility, Web content, Natural language processing, Arabic language

Paper type Research paper

1. Introduction

Twitter is a micro-blogging service that provides a light-weight, easy form of communication which enables users to broadcast and share information about their activities, opinions and status (Java *et al.*, 2007). In recent years Twitter has grown vastly in terms of users and content. According to Alexa.com, more than 73 million global internet users visit Twitter daily.

Twitter and similar micro-blogging services are considered a source for up-to-date information about what is happening in the world (Grinev *et al.*, 2009).Their ability to deliver data to interested users over multiple delivery channels (Krishnamurthy *et al.*, 2008)

The authors are thankful to Professor AbdulMalik S. Al-Salman, for his valuable and thoughtful feedback.

International Journal of Web Information Systems
Vol. 7 No. 2, 2011
pp. 130-151

1744-0084
DOI 10.1108/17440081111141772

have marked them as a potential place for rumors and gossips. Thus, questioning their content credibility.

Content credibility has its roots in many research disciplines including: psychology, communication, information science, marketing/ecommerce (Kwon *et al.*, 2009), human-computer interaction (HCI) (Lazar *et al.*, 2007), information retrieval and web engineering (Kamthan, 2008).

During the past decade many studies on web information credibility were carried out to find the theoretical framework for evaluating web sites and to define elements that aid in designing credible web sites, as it is the case with the web credibility project at Stanford University (Rieh and Danielson, 2007).

Recently, the need to evaluate electronic information credibility opens new research area for computer science researchers to define automatic measures for evaluating web sites' contents or help users in their judgments. Most of the available research aim to measure the credibility of text content on the web, especially those with user generated content such as weblogs (Weerkamp and De Rijke, 2008; Juffinger *et al.*, 2009), discussion forums (Weimer *et al.*, 2007; Wanas *et al.*, 2008) and Wikipedia (Lopes and Carriço, 2008), in addition to other types of web sites, e.g. portals (Kidawara, 2008; Akamine *et al.*, 2008) and in the semantic web (Kaczmarek, 2008). Multimedia content, such as video and audio, has also been a concern of credibility research; however these studies are still in their initial stages (Nakamura *et al.*, 2008; Tsagkias *et al.*, 2008).

When it comes to Twitter as a service that disseminates text content, we can find that the studies on Twitter credibility have not started yet. However, there are many studies on understanding Twitter usage and communities, e.g. (Krishnamurthy *et al.*, 2008; Honey and Herring, 2009; Java *et al.*, 2007), sentiment analysis (Parikh and Movassate, 2009), Twitter user re-identification (Narayanan and Shmatikov, 2009), Twitter News search (Grinev *et al.*, 2009; Sankaranarayanan *et al.*, 2009) and Twitter recommendation (Phelan *et al.*, 2009).

In this paper, we will focus on automatically measuring the credibility of Arabic News content published in Twitter. The choice of the News domain is attributed to the new changes in the social and political climate in the Arab region; which caused the Arabic content to draw attention worldwide.

This study can be considered a first attempt to explore the credibility of Twitter content targeted for the Arabic language. Our study builds on features used in previous credibility studies in addition to custom compiled features specific for Twitter. We also run experiments to evaluate the effectiveness of using different features and classification techniques on our dataset.

Therefore, our paper is structured as follows: first, we present and discuss the theoretical background and previous work related to our project theme. Next, we present the design and implementation of our system, and the candidate features for evaluation. Then, we evaluate our proposed system by conducting three experiments and discuss their results. Finally, we conclude the paper by highlighting our system limitation and point out further research directions.

2. Theoretical background

Credibility research began in 1950s in the fields of communication and psychology (Tseng and Fogg, 1999). The simple definition of credibility is: believability. Credible people are believable people; credible information is believable information

(Tseng and Fogg, 1999). However, there is no clear definition of the term credibility until now (Hilligoss and Rieh, 2008).

In 2001, Fogg *et al.* (2001) defined the concept of credibility that is related to web sites and noticed that credibility does not depend on the object, person or piece of information; it is a perceived quality resulting from evaluating multiple dimensions simultaneously. That is why there are many empirical studies discussing the perception of credibility (Iding *et al.*, 2008). As a result, credibility has different dimensions and factors that affect web credibility evaluation. But the major component perceived of web site credibility is from trustworthiness and expertise. Trustworthiness means the perceived goodness and morality of the source while expertise means the perceived knowledge and skill of the source (Fogg *et al.*, 2001).

Credibility has its roots in many research disciplines with different approaches and goals including: psychology, communication, information science, marketing/ecommerce (Kwon *et al.*, 2009), HCI (Tseng and Fogg, 1999), information retrieval and web engineering (Kamthan, 2008). In this research we will focus specifically on information credibility on the web.

There are many studies during the past decade on web information credibility dedicated to find the theoretical framework for evaluating web sites and to define elements which aid in designing web sites that appear to be more credible, as it is the case with the web credibility project at Stanford University (Rieh and Danielson, 2007). In order to design a credible web site, Stanford University developed guidelines for web credibility (http://credibility.stanford.edu/guidelines/index.html). The guidelines include: accuracy of information, clear organization of a web site, trustworthy of people that stand behind the web site, easy contact information, usable web site, frequently updated content, restraint any promotional content and error free web site.

However, in recent years there has been an attempt to enforce credibility within the development process of web applications to overcome the limitations of previous theoretical and empirical studies (Fogg *et al.*, 2001; Fogg, 2003) which lack ensuring credibility only by guidelines with no feasibility meaning (Kamthan, 2008). As a result, Kamthan defines the concept of credibility engineering based on the consumer interaction with a web application as "the discipline of ensuring that a Web application will be perceived as credible by its stakeholders and doing so throughout the life cycle of the web application". He also proposed a framework for active, surface and experienced credibility, toward systematic approach of web credibility (Kamthan, 2008). This framework considers three dimensions of credibility in web applications which are taken from managerial, societal, and technical viewpoints (Kamthan, 2008). Furthermore, Kamthan called for the understanding of credibility concept to be considered a mandatory non-functional requirement expressed earlier and subsequently attended throughout the web application development process.

3. Related work

Text content can be found in different genres of web sites especially those with user generated content such as weblogs and micro-blogging services. In fact, weblogs and micro-blogging are considered the same (both contain an online personal journal with reflections, comments, and often hyperlinks provided by the writer Mishne (2007)). The major difference between the two is in the post length. In weblogs, the length of a post is unlimited while in micro-blogging services the limit is usually 140 characters.

Weblogs have a broad area for exploration in terms of credibility assessment due to the high level of self-disclosure of information that might reveal trustworthiness and expertise by bloggers and availability of audience evaluations (Rubin and Liddy, 2006).

Weblogs credibility studies started by House (2004). House made some observations about the importance of credibility evaluation in Weblogs due to their high degree of self-disclosure. These observations were considered afterwards by Rubin and Liddy to develop their own Weblogs credibility assessment framework. Thus, the actual opening study of the measurements of Weblogs credibility automatically by Rubin and Liddy (2006).

Rubin and Liddy suggested the use of a framework of Weblogs credibility assessment based on four factors: Blogger expertise, blogger's trustworthiness and value system, information quality and appeal and triggers of a personal nature. They derived these factors from credibility theory studies, Stanford survey studies (Tseng and Fogg, 1999; Fogg *et al.*, 2001), House observations and their previous experience in weblogs.

To determine the credibility features for weblogs, Ulicny and Baclawski (2007) provided a way to select features that clearly distinguish between a set of weblogs: those considered as previously credible, from those considered ordinary weblogs. Therefore, they suggested some features related to blogger profile and its writing. The features include the following: Blogger profile Full Names, Affiliation, Comments Unquoted Content, Links to News Sources, average number of sentences, average number of Paragraphs and average Readability. They also concluded that features related to writing style alone do not distinguish credible bloggers.

"More credible blog posts are preferred by searcher", this hypothesis was implemented and tested by Weerkamp and De Rijke (2008). They incorporated textual credibility indicators into the topical blog post retrieval. Their credibility features were selected from Rubin and Liddy weblogs credibility framework. They selected features that are textual in nature and not related to blogger's identity. Also they selected features that can be estimated automatically from their test collection and can be estimated with state-of-the-art language technology. They grouped their indicators into: blog level and post level, in addition to: topic dependent and independent. Then they performed their experiments on the text retrieve conference blog track test set and showed that both groups of credibility features significantly improved retrieval effectiveness. The best performance of a single feature was the post length and comments, while within group of features it was the topic dependent post level. Furthermore, the best performance was achieved when combining all the features. Their future work will be to use topic dependent, by analyzing blogger profile with stated competencies indicator and using reading level measures (e.g., Flesch-Kincaid) for the literary appeal feature.

Moreover, Conrad *et al.* (2008) have shown that using singular metric specifically "inlink counts" is not effective to measure authority and presumed credibility of weblogs. They discussed credibility of the author especially in Professional information providers in the legal and finance domain. They considered authority as a proxy of credibility. As a result they identified ten classes of features for measuring authority. These features include: activity level and popularity as two baseline methods, as well as different set of main features. In addition statistical features were derived from the author writing style as a supplementing dimension.

In order to distinguish between gossip and credible blogs in the News domain (Juffinger *et al.*, 2009) improved their blog analysis system in Austria Press Agency to rank only German blogs according to their credibility. They showed that credibility of blogs is derived mainly from its content. So, they measured credibility based on two new dimensions: quantity structure and similarity with verified content. Their credibility ranking process passed through two phases. The first phase was the quantity structure, by computing the correlation between blog posts and news articles over time, strong correlation reveals that blogs are not spammed. In the similarity phase, they performed some natural language processing (NLP) tasks to each news article and blog post, then computed the centroid cosine similarity that takes into account nouns/verbs and objectives. Finally, their system assigned one credibility rating to each blog from three levels: little, average or high credible. The results of the ranked blogs were verified by some human experts.

We also noticed from previous research specific to Twitter that there are some studies in the news domain, which considered Twitter as an effective news media to spread news much faster than conventional news media. For example, Grinev *et al.* (2009) developed a Twitter search system which is called TweetSieve. The system describes news events that are given as query by showing the time period of this news event and the best matched tweets. They used Jaccard similarity measure in order to cluster tweets by comparing them to each other and finally finding a central tweet as the main representative on each cluster. Yet, they did not do any evaluation of their system to evaluate its performance.

Sankaranarayanan *et al.* (2009), on the other hand, presented a method to automatically identify important news events from Twitter data in their TwitterStand system. They discarded junk news, or non-news tweets by using naive Bayes classifier, which is trained on a training set of tweets marked as news or junk. Then they applied online clustering algorithm to automatically cluster tweets into the corresponding news stories, based both on their content and time. For clustering content, they used cosine similarity measure with term frequency inverse document frequency (TFIDF) weight to assign each tweet to the appropriate cluster. For the time period, they used Gaussian distance between each tweet and the time centriod of a cluster. The important work in this step is the reduction of noise by marking the clusters that do not contain any tweets as inactive. Finally, they extracted a location for each cluster of tweets, either by analyzing its content using NLP techniques such as part of speech tagging (POS) and name entity recognition, or by extracting location information from tweets meta data directly. Some of the techniques mentioned in this research have inspired our framework design decisions.

Finally, Phelan *et al.* (2009) proposed a novel news recommendation system based on Twitter data as a source of current and topical news. They used Twitter data, either from public timeline or only from friend timeline, for ranking articles from a collection of RSS feeds entered by a user. For the ranking algorithm, they used TFIDF measure to compute the co-occurrence of terms between RSS feeds and Twitter data. Their preliminary evaluation on ten participants reached a conclusion that users can benefit from the recommendations derived from Twitter data.

From the previous credibility research we can observe that there are different features used to measure weblogs credibility. The difference in feature selection can be attributed either to the research domain or to its objective. For example, features used to determine

credibility of News are different from the legal domain which focuses on professional expertise (authority) rather than trustworthiness of the News. However, most research agree to measure credibility in two aspects: author and content. For author features the most used feature was the user profile, while for content features it was verifying the authority of content by linking to authoritative sources or/and computing the content similarity with authoritative source. Moreover, most of the used content features depend on linguistic characteristics, i.e. stemming and POS. These studies indicate that language dependent features are most important to asses.

4. Proposed system

Based on our previous research in weblog credibility, we can distill a set of features applicable for evaluating Twitter credibility. We will also propose new features that are specific to Twitter. Thus, this section highlights the candidate credibility features used to evaluate Twitter content as well as the proposed system architecture.

4.1 Twitter features

Twitter has its own unique characteristics that are different from other web sites. It can be seen as a set of tweets, each of which is created by one user and seen by other friends or followers. A Twitter *user* is a person or a system that posts a tweet or a message on Twitter, and a *tweet* is a short message from a user to its set of followers (Sankaranarayanan *et al.*, 2009). Each tweet and user profile has a set of features that can be defined and described in Table I (for Tweets) and Table II (for users). Given the set of features mentioned in the previous two tables as well as the features experimented in previous research studies; five features (three for Twitter content and two for user profile) were selected for further investigation. These features are listed in Table III. Notice that the most important feature for Twitter content credibility is the similarity with verified content. This feature has been used before by (Juffinger *et al.*, 2009) and it obtained successful results in measuring weblogs credibility.

Feature	Description
Linking	A shorten URL that direct a user to external resources using services that shorten links to under 10-20 characters in length. Some of these services are: http://tinyurl.com http://bit.ly
Addressivity	Directs a message to another person by showing the presence of @ at the beginning, e.g.: @ + username + message[a]
Retweet	A way to spread information through the Twitter network by showing: the @ not at the beginning RT with @ at the beginning Username alone, e.g.: RT + @ + username + message username + message[b]
HashTagging	It is similar to other web tags- it helps add tweets to a category Hashtags have the "hash" or "pound" symbol (#) preceding the tag[c]

Notes: [a]http://help.Twitter.com/forums/10711/entries/14023; [b]http://help.Twitter.com/forums/10711/entries/77606; [c]http://help.Twitter.com/forums/10711/entries/49309

Table I.
Features of Tweets' content

Feature	Description
Username	A user handle in Twitter
User picture	User image
Real name	The full user name
Location	The user location (country, city, etc)
Web site link	The URL of the user web site
Bio	A brief biography of the user
Number of tweets	The total number of tweets sent by the user
Followers	Followers are people who receive a user tweets. If a user follows someone, he/she becomes a follower – the user will then receive others tweets in his/her homepage. If someone follows a user, they will be the user follower – they'll receive the user tweets in their home page, phone, or any application.[a]
Friends	User follow updates from other members who are added as a friend
Verified account[b]	Determine which accounts we know are "real" and authentic. This means that Twitter has been in contact with the account holder and verified that it is authentic

Table II.
Twitter user profile features

Notes: Most user information can be found at the right top pane of the user interface; [a]http://help.Twitter.com/forums/10711/entries/14019, [b]http://Twitter.com/help/verified

Type	List of features	Used by
Content	Similarity with verified content	Juffinger *et al.* (2009)
	Inappropriate words	Sankaranarayanan *et al.* (2009)
	Linking to authoritative /credible news sources	Ulicny and Baclawski (2007)
Author (or user)	Is the user verified?	Ulicny and Baclawski (2007)
	What is the overall degree of a Twitter user? (explained later)	

Table III.
Nominated Twitter credibility features

4.2 System architecture

The proposed system architecture (Figure 1) consists of four main components: text preprocessing, features extraction and computation, credibility calculation and credibility assignment & ranking.

4.2.1 Text preprocessing. For the preprocessing steps, Twitter tweets are normalized; stop words are removed, and words in a tweet are POS tagged and stemmed. However, before applying the stemming algorithm to News and tweet texts, we need to perform POS in order to measure the similarity based on different POS tags. This is because nouns generally cover the thematic information in a news story while verbs indicate the associations with a topic (Juffinger *et al.*, 2009).

Figure 2 shows the main components of the text preprocessing module.

The input consists of relevant tweets and news articles for a specific query at the same time period. The preprocessing normalizes the tweet texts, removes stop words, performs stemming and POS. Then each document, either tweet or news article, is represented as a bag of words (BOW).

Two Arabic NLP tools were used for the preprocessing phase. Khojah Stemmer[1] was used to remove prefixes, suffixes and Arabic stop words, and return the words stem. The Stanford part of speech tagger[2] was used to assign part of speech tags to each word in a tweet.

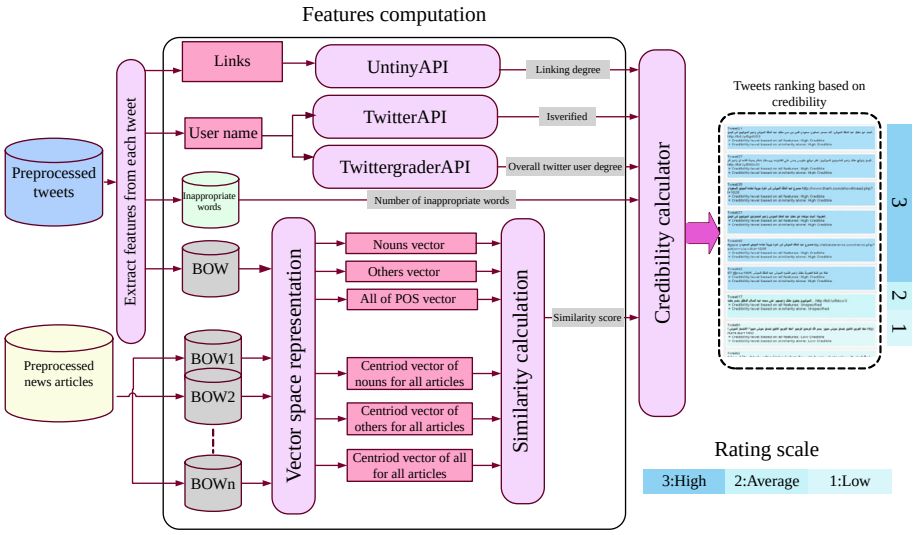

Figure 1.
System architecture

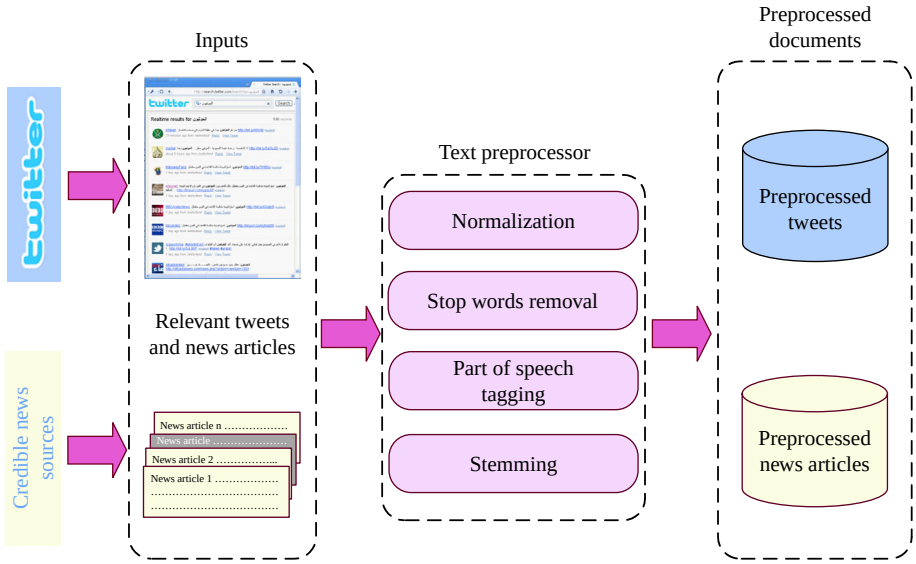

Figure 2.
Text preprocessing components

When POS is finished, each document, either tweets or news articles, are represented as a BOW. BOW representation contains the list of distinct terms in a document with the corresponding absolute term frequency (ATF).

4.2.2 Features extraction and computation. The features computation module uses the cosine similarity measure to compute the value of the similarity with verified content feature. As for the other features, we used available web services to extract them. Next, we describe in further details the set of features designated for both the content and the author.

4.2.2.1 Content features. Tweet content consists of three features which are: similarity with verified content, linking to authoritative/credible News Sources and the appearance of inappropriate words (IW):

(1) *Similarity with verified content.* To calculate the similarity with verified content value, we followed the following systematic steps (Figure 3).

The *first step* is to generate a list of distinct terms from news articles. Then we compute the document frequency $\boldsymbol{df_i}$ for each term $\boldsymbol{i}$ in the list.

The *second step* is to compute the inverse document frequency $\boldsymbol{idf_i}$ for each term $\boldsymbol{i}$. The inverse document frequency is calculated using equation (1) (Harrag *et al.*, 2009):

$$\boldsymbol{IDF}_i = \log_2 \frac{\boldsymbol{numberofdocuments}(N)}{\boldsymbol{DF}_i} + 1 \tag{1}$$

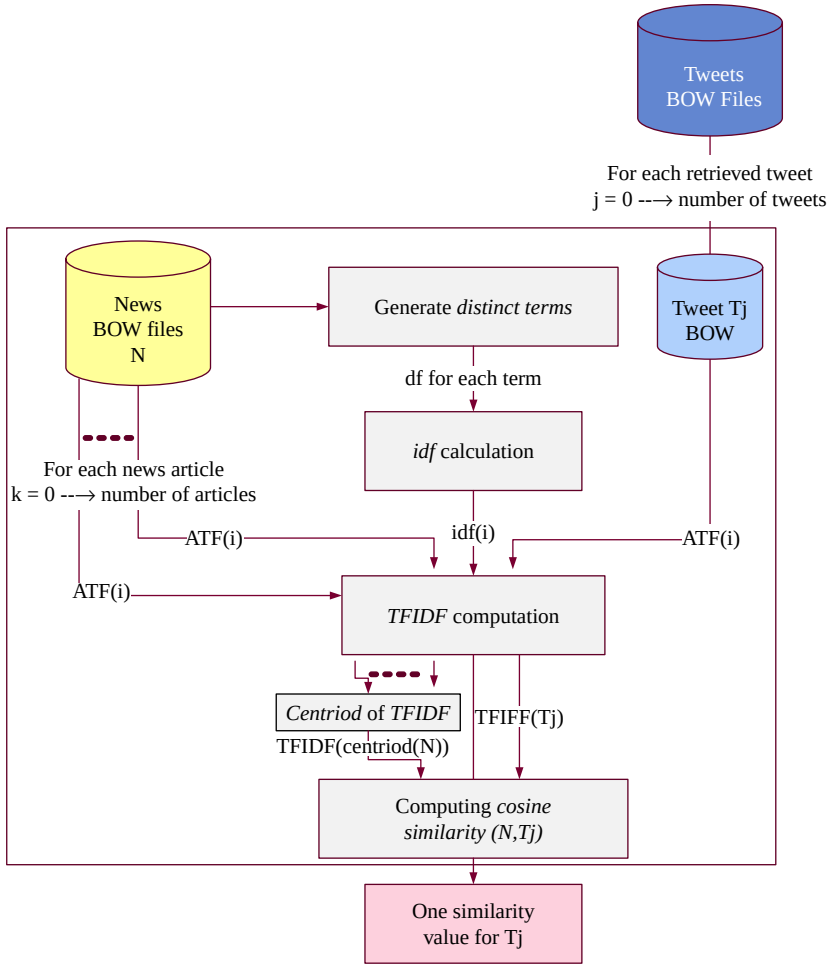

Figure 3.
Steps for computing the similarity value

However, if one or more terms are not available in the tweet, this will lead to dividing by a zero document frequency. To resolve this issue, we use an alternative equation (2) only for this situation by setting the document frequency to 1 and increasing the number of documents N by one. As a result, equation (1) will be changed to equation (2):

$$IDF_i = \log_2 \frac{numberofdocuments(N+1)}{1} + 1 \quad (2)$$

i: refers to a tweet term number that is not available on the list.

The *Third step* is to compute the weight of the TFIDF for each term i in document j as in equation (3) (Harrag *et al.*, 2009):

$$TFIDF_{ij} = TF_{ij} \times IDF_i \quad (3)$$

The term frequency tf_{ij} has many approaches, so we selected the frequently used one which equals to ATF.

The *fourth step* is to compute the centriod vector for all news articles. Thus, for each term i we calculated the weight of $TFIDF$ using equation (4):

$$TFIDF_i = \sum_{j=0}^{N} \frac{TFIDF_{ij}}{N} \quad (4)$$

Finally, we computed the three similarity values for each tweet based on the different POS between each tweet **Tj** and the Centroid of the news Articles **C**. We used a cosine similarity measure as defined in equation (5):

$$sim(Tj,\ Centroid(N)) = \frac{\sum_{termicTNC} TFIDF_{ij} \times TFIDF_{iC}}{\sqrt{\sum_{termicT} TFIDF_i^2 \times \sum_{termicC} TFIDF_i^2}} \quad (5)$$

(2) *Inappropriate words.* This feature is defined as the presence of words that are on the list of IW. This list was compiled manually after examining Twitter posts.

(3) *Linking to authoritative/credible News sources.* This feature means the inclusion of a link that refers to an authoritative source. First, we need to restore the short URL to its original form using the Untiny.com web service. Then we check the type of the original web site and give it a degree (High, Medium or Low credibility) based on its popularity and credibility. This degree is assigned by a human expert in the Political Science domain.

4.2.2.2 Author features. Author features consist of two factors: first is the Twitter user verified? and second the overall user rating:

(1) *Is verified.* This feature is already provided by Twitter.com to its users and it has two values: true or false.

(2) *TwitterGrader.com value.* Twittergrader.com is a service to retrieve the grade of any Twitter user via its username. Basically it measures the power, reach and authority of a twitter account based on a set of factors which include: number of followers, power of followers, updates, update recency,

follower/following ratio and engagement. We obtained a trail period of the service API Key to use it in our system.

4.2.2.3 Compiled credibility features. At the end, we have values for the list of nominated features. The features' values are used as parameters in our credibility formula in order to calculate the credibility score. Table IV shows the list of complied features along with their values.

4.2.3 Credibility calculation. The credibility calculator receives the results of the feature extraction module as an input to its credibility formula then calculates the final credibility score for each tweet.

Two approaches were used to calculate the final credibility score and classify tweets into three credibility levels, namely:

(1) similarity with verified content; and

(2) similarity with verified content in addition to extra features.

4.2.3.1 First approach. The *first approach* is to classify tweets into three credibility levels based only on the similarity with verified content feature. Thus, we implemented automatic thresholding using similarity values with different POS as defined by (Juffinger *et al.*, 2009).

The thresholds were computed based on different POS tags. In the first equation (6), the similarity values are computed based on nouns only. The same procedure is applied to the similarity values computed based on all POS, shown in equations (7 and 8):

$$\boldsymbol{T_1} = [\mathbf{0}, \mathbf{min}(\boldsymbol{simOfNoun}] + \mathbf{0.5} \times (\mathbf{max}(\boldsymbol{simOfNoun}) - \mathbf{min}(\boldsymbol{simOfNoun}))] \quad (6)$$

$$\boldsymbol{T_2} = [\mathbf{0}, \mathbf{min}(\boldsymbol{simOfOther}] + \mathbf{0.5} \times (\mathbf{max}(\boldsymbol{simOfOther}) - \mathbf{min}(\boldsymbol{simOfOther}))] \quad (7)$$

$$\boldsymbol{T_3} = [\mathbf{0}, \mathbf{min}(\boldsymbol{simOfAll}] + \mathbf{0.5} \times (\mathbf{max}(\boldsymbol{simOfAll}) - \mathbf{min}(\boldsymbol{simOfAll})), \mathbf{1}] \quad (8)$$

The intervals were determined after many experiments conducted by Juffinger *et al.* which showed that low credibility happens either when posts are dealing with "a different topic (from nouns) or are in a completely different association with the topic (from verbs and adjectives)".

Therefore, we used the intervals (T1 and T3) proposed by (Juffinger *et al.*, 2009) and added a new interval T2, which considers all POS except nouns. The rationale behind adding this new interval is because Arabic tweets are short (max 70 characters),

Table IV. Compiled credibility features

Feature	Inputs to credibility formula
Similarity with verified content (S)	S_n: Similarity value based on nouns S_o: Similarity value based on other POS tags excluding nouns S_{all}: Similarity values based on all words without restriction of POS tags
Inappropriate words (IW)	IW: Number of inappropriate words
Linking to authoritative /credible News sources (L)	L: Degree of news web site credibility
Is verified (V)	V: 1 if true and 0 if false
TwitterGrader.com degree (G)	G: Degree of Twitter user in percentage

which sometimes might lead to a limited number of POS tags. So to ensure that other POS are considered within the interval of nouns, we added this extra interval.

Next we followed the following three steps to categorize the tweets into credibility levels:

- The *first step* is based on similarity values of nouns. All tweets in the interval of threshold T_1 are assigned level 1 (low credibility).
- The *second step* is based on similarity values of other POS. All tweets in the interval of threshold T_2 are assigned level 1 (low credibility) as well.
- The *third step* is based on similarity values of all POS tags. Tweets in the interval of threshold T_3 are assigned level 3 (high credibility).
- Finally, the remaining tweets that do not belong to any similarity intervals are assigned level 2 as "average" credibility.

4.2.3.2 Second approach. The *second approach* is to use extra features to compute the credibility of Twitter tweets, as shown in equation (9):

$$\boldsymbol{Credibility\ \ Score = 0.6S + 0.2IW + 0.1L + 0.1A} \tag{9}$$

Where:

S refers to the similarity value computed previously, i.e. (S_n, S_o, S_{all}).

IW refers to Inappropriate Words.

L refers to Linking to authoritative source.

A refers to Author feature, it is computed from two variables: is Verified (V) and TwitterGrader degree (G).

Each feature in the equation was given an experimental weight based on its importance and frequency of appearance in a tweet. From our own observations, we found that the linking feature appears more often in a news tweet; however the appearance of IW will affect the tweet credibility, so we setup the total weighting of these two features to 30 percent. On the other hand, we weighted the author feature to 10 percent because we did not know exactly how the algorithm behind the authentication service works. As a result, 90 percent of our final formulated equation focused on the content rather than the author, which makes it subject to further refinement as we have suggested in section 6.

4.2.4 Credibility assignment and ranking. Finally, the system assigns credibility levels to each tweet based on the computed credibility score. It rates the tweets based on their credibility level into three levels: high, low and average credibility.

5. System evaluation

Three experiments were conducted to evaluate the performance of our system, that is: similarity with verified content, similarity with extra set of features, as well as benchmarking the system results against human evaluation.

5.1 Dataset

In order to evaluate our system, we compiled a data set that consists of tweets and news articles. Our data set was constructed by applying a set of queries around two hot political topics suggested by political experts, these topics were likely susceptible to

gossip at the given time frame of the experiment. The queries were submitted to Twitter and two authoritative news sources on the web, which are: the Saudi Press Agency (SPA) and Aljazeera.net, in addition to different news sources from Google News.

Data were collected for around two weeks (from December 27, 2009 to January 6, 2010); the result was approximately 600 tweets and 179 news articles. Google reader was used to help in collecting and indexing the data set. The details of the collected data are given in Table V.

5.2 Experiments' results

We conducted three experiments to evaluate the result of our system. To the best of our knowledge, there was no standard evaluation method for measuring the credibility of Twitter to benchmark the results of our system against. Thus, we used the common evaluation methods used in information retrieval domain which include: the human evaluation in addition to the traditional measures of Precision, Recall and F-Measure.

5.2.1 First experiment. The aim of the first experiment is to compare the result of using all set of features for measuring Twitter credibility opposite to using only the feature of similarity with the verified content. For this experiment we used 268 tweets and 17 representative news articles from the data set of topic 2 (Yeman and Houthi).

5.2.1.1 Results of the first approach: similarity with the verified content. For the similarity with verified content feature, we first calculated (for the given data set) the three different similarity thresholds mentioned previously, then we used the computed thresholds (Table VI) to classify the tweets.

Results obtained from this experiment showed that 42 percent of the classified tweets using similarity values for nouns and 40 percent using similarity values of Other POS were assigned "Low Credibility". However, 16 percent of the tweets classified using similarity values of all POS were assigned "High Credibility".

5.2.1.2 Results of the second approach: similarity with extra features. For the second approach, we applied equation (9) which uses the similarity feature plus a set of proposed features to calculate the final credibility score. In this approach, the results show a change in the credibility rating for 12 percent of the tweets from level 1 or 3

Topic 1	إيران	[Iran]
Number of tweets		330
Number of news articles		2 from SPA & Aljazeera
Topic 2	اليمن والحوثيون	[Yeman and Houthi]
Number of tweets		268
Number of news articles		46 from SPA 16 from Aljazeera 97 from Google News

Table V.
Data set collection

Threshold	Range
Threshold based on similarity of Noun POS (T1)	[0,0.22805]
Threshold based on similarity of Other POS (T2)	[0,0.2118]
Threshold based on similarity of ALL POS (T3)	[0.2163,1]

Table VI.
Computed thresholds range

(Low and High credibility) to level 2 ("Average Credibility") and leaving the rest of the tweets rating unchanged.

5.2.1.3 Discussion of the two approaches. Figure 4 shows the results of the three credibility levels with the corresponding number of tweets rated in the two approaches. We can observe from the obtained results that the first approach (similarity feature with verified content) was able to assign tweets to only two credibility levels (Low and High), while in the second approach (similarity plus extra features) was able to assign the tweets to all three levels of credibility.

It can be seen also from Figure 4 that level 2 (average credibility) rating is low in the second approach, or does not appear in the first approach. This variation in rating might be attributed to: first, the subtle difference in the classification range (hence Table VI) or second, the effect that the extra features have added to the credibility formula.

To address the first issue, the misclassification of level 2 based on the similarity feature alone might be due to the actual similarity value that ranged from 0.1691 to 0.20810 for 95 percent of the tweets. So, for the similarity values of nouns most of the tweets are below the nouns similarity threshold and as a result, most of them will be assigned "Low Credibility", and so on for the other similarity threshold.

We may also notice from Table VI that most of the similarity values (T1 and T2) are small and ranged from 0 to less than 0.3. These small values are due to the comparison of short documents, i.e. tweets against long news documents. In this case, the number of similar terms between the two documents is limited to the number of terms on each tweet, which are around 10 terms on average. As a result, when comparing the terms in the news articles, which are usually five or more times greater than the tweet terms, the similarity result will be smaller. In fact, we can overcome this problem by dividing the news articles into sentences and then computing the similarity between a tweet and each sentence. Finally the average rate of these similarity values will be calculated and used. This technique is similar to (Nagura *et al.*, 2006) for rating news documents on the web.

In contrast, "average credibility" rating has appeared while using the second approach (similarity plus extra features). In this approach, our system raised the levels of some tweets to "average credibility", while it leaved most of the tweets at the same level, either high or low. This can be the effect of one or more used extra features.

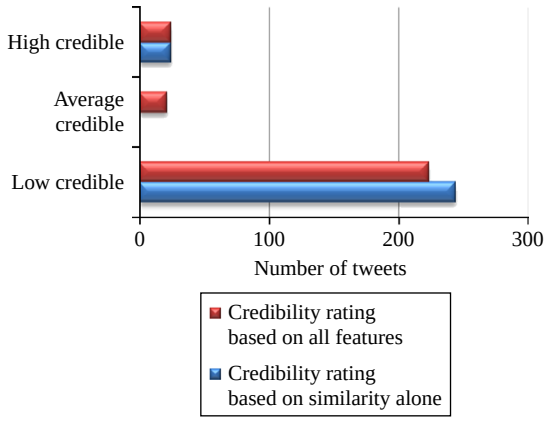

Figure 4.
Comparison between the credibility ratings of the two approaches

By going back to the set of extra features, we start with the Author Features (A) that is defined by V and G. We can see that feature V (isVerified) has a zero values for all Twitter users on our data set. This may be attributed to the actual lack of verified Twitter users in our data set; however the main reason can be traced back to the minimum number of Twitter users that holds this verification from Twitter. Thus, this feature does not affect the credibility score at all, but it may have more impact in the future when considering verified users by Twitter.

For the second parameter of the author feature that is the user overall degree (G), the values returned by Twittergrader.com ranged between 40.68 and 42.73. These values are not small and they might indicate an effect on the author degree and finally on the credibility score.

For the Linking feature (L), we can find that most tweets include a link to a web resource; however the main impact of this feature is when it is linked to an authoritative source that has a high credibility degree.

The last feature is the appearance of IW, which is almost zero in our data set. This may refer to the little use of IW on news tweets or may be due to the limited number of words that we have collected.

As a result, we can say that the most significant feature which has affected the credibility rating on our data set is either the user overall degree (G) or the linking feature (L). Nevertheless, we can consider the linking feature having the most impact on credibility, since we do not know the Twittergrader.com degree algorithm. This does not mean ignoring the effect of other features but still this issue needs to be further investigated to expose the set of features that will affect Twitter credibility.

5.2.2 Second experiment. The second experiment was intended to show the accuracy of our results using the traditional measures of Precision, Recall and F-Measure.

The *Precision* is calculated as the number of tweets automatically classified into one of the credibility levels by our System and also classified to the same level by the human experts. Thus, the Precision measure for the credibility level (L) is calculated as follows:

$$\text{Precision (L)} = \frac{\textbf{\textit{Human classified to L} } \cap \textbf{\textit{ our system classified to L}}}{\textbf{\textit{our system classified to L}}}$$

The Recall is calculated as the number of tweets classified to one of the credibility levels by human experts and also classified automatically to the same level by our system. The Recall for the credibility level (L) is expressed as follows:

$$\text{Recall (L)} = \frac{\textbf{\textit{Human classified to L} } \cap \textbf{\textit{ our system classified to L}}}{\textbf{\textit{Human classified to L}}}$$

The F- Measure is calculated using the following equation:

$$\text{F} - \text{Measure } (L) = 2 \times (\text{Precision} \times \text{Recall}/\text{Precision} + \text{Recall})$$

For this experiment, we manually selected from our data set 29 tweets and four news articles. From topic 1 we selected 9 tweets and 2 news articles while from topic 2 we selected 20 tweets and 2 news articles. We sent the data set for rating to three human experts in the political science domain.

Tables VII–IX present the Precision, Recall and F-Measure of our system with reference to evaluator1, evalautor2 and evalautor3 results, respectively.

It appears from the F-measures' results (Figure 5) that the similarity alone approach outperforms the similarity with all features in the "low credibility" rating for the three evaluators. However, in the "high credibility" rating, we can see that the two approaches have the same score. Going back to the data we can find that the precision and recall were the same for the two evaluators which lead to this agreement in F-Measures.

Also, there is a lack of "average credibility" rating even though some evaluators have distinguished some tweets as being from this type. This might be because our system has low number of rated tweets in this category especially in the second approach, which does not overlap with the evaluators' rating.

Table VII.
The precision, recall and F-measure of our system with respect to evaluator 1

	Low		Average		High		F-measure	
Approach	P	R	P	R	P	R	F-low	F-high
Similarity alone	0.73	0.76	N/A	0	0.29	0.4	0.75	0.34
All features	0.76	0.62	0	0	0.29	0.4	0.68	0.34

Table VIII.
The precision, recall and F-measure of our system with respect to evaluator 2

	Low		Average		High		F-measure
Approach	P	R	P	R	P	R	F-low
Similarity alone	0.91	0.8	N/A	0	0	N/A	0.85
All features	0.88	0.6	0	0	0	N/A	0.71

Table IX.
The precision, recall and F-measure of our system with respect to evaluator 3

	Low		Average		High		F-measure	
Approach	P	R	P	R	P	R	F-low	F-high
Similarity alone	0.82	0.78	N/A	N/A	0.29	0.33	0.8	0.31
All features	0.76	0.57	0	N/A	0.29	0.33	0.65	0.31

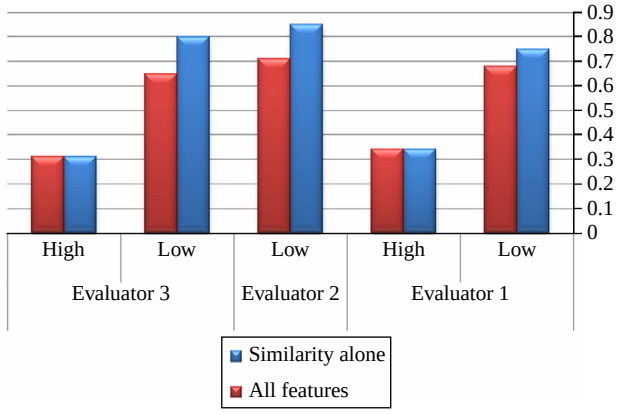

Figure 5.
The F-Measures of the two approaches (similarity alone and similarity with all features) for the three evaluators

Finally, all evaluators have a high F-measure value for "low credibility" contrary to "average" and "high". This can be attributed to the fact that most Arabic news distributed via Twitter has a low degree of credibility from the experts' point of view.

5.2.3 Third experiment. The final experiment was conducted to show the accuracy of our system based on the human experts' evaluations. We used the same data set of our second experiment; however, we depended only on two of the human experts, namely evaluator 2 and 3. This is because the kappa inter-rater agreement value of the two human evaluators was 0.6, which reflects a good agreement between the two evaluators.

Table X and Figure 6 show the evaluation conducted by the two experts opposed to our system results and for the same data set.

From the result we can see that each evaluator has assigned only two levels of credibility either low and high or low and average. There are many interpretations for this disagreement which include: individual difference, bias, previous experience, etc. which are beyond our paper scope and subject for further investigation.

On the other hand, the "low credibility" rating is the most credibility level agreed by the evaluators and our system.

6. Discussion

From the three experiments we can conclude that the first approach (similarity with verified content) has proven its effectiveness in evaluating the credibility of Twitter tweets.

Table X.
The percent of each credibility level of our system and the two evaluators

	Eva.2	Eva.3	AVG	Similarity alone	All features
Low	86	79	83	76	59
Average	14	0	7	0	17
High	0	21	11	24	24

Note: Values in percentage

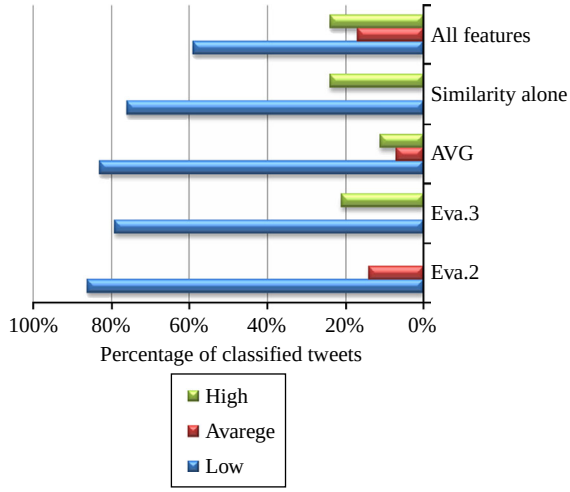

Figure 6.
The percentage of each credibility level on our system against the two evaluators

This result is similar to (Juffinger *et al.*, 2009), where they used similarity with verified content to evaluate the credibility of Weblog posts. However, when using this feature alone, our system rated the tweets to only two credibility levels: Low and High, and could not distinguish "average credibility". This problem might come from the values of the similarity thresholds (T1 and T3), which are based on (Juffinger *et al.*, 2009) without deriving it experimentally based on our dataset. Also the increase of the "low credibility" rating might be the reason for the high precision of the similarity feature than the other cases seen in the previous experiments.

In contrast, when using the second approach (similarity with verified content plus extra features), the system was able to recognize "average credibility" tweets. Based on our observations the major feature affected the second approach was the linking feature. However, the weighting of the features in the second approach needs to be further examined and investigated.

Given the previous discussion we can notice that the nature of Twitter as a micro-blogging service, which consist of short tweets that do not exceed 140 characters in Latin-scripts and 70 characters in Arabic, have affected the categorization of credibility levels. In fact, the similarity algorithm used in determining the credibility of tweets was affected directly by this limitation. Which led to classifying tweets into two classes "low" and "high", however, when we did not rely only on the similarity algorithm alone, i.e. by adding extra features devised from the content of a tweet, the "average" credibility became more apparent.

This drives us to think about studying other Twitter features such as hashtags, re-tweets, emoticons, redirects, etc. and observe how they affect the credibility of tweets. We also need further investigation on how to measure the credibility of Twitter users instead of using unknown algorithms from web services tools. One suggestion might be using Twitter user profile and other social measures such as influential ratio, user activity, etc. to de-anonymize users by using other social web sites such as Facebook or Flicker as an indicator for credibility of Twitter users. This idea of de-anonymizing social networks is done by (Narayanan and Shmatikov, 2009).

7. Conclusion

In this paper, we presented a system for measuring the credibility of Arabic Twitter news using two approaches: similarity with verified content and similarity with extra features. Our experimental results showed that our system has a good performance when using similarity alone, compared to the set of all features.

Despite the small effect of the other features, previously suggested features for calculating and measuring Twitter credibility provided a reasonable precision degree. In fact, the linking feature has the most impact than other features. This does not mean ignoring the effect of the other features, yet further experiments are needed to justify this finding.

While building our system and conducting our experiments we noticed that even though Twitter is considered a genre of weblogs, yet, not all features that are applicable for weblogs where used in our system. This is because Twitter tweets are very short, the infrastructure of Twitter is more inter-related than weblogs, and Twitter has its own features that do not exist in weblogs such as followers, re-tweets, and hashtags.

When we tried using the measures of weblogs, we were enforced to modify them to cope with the nature of Twitter tweets, as illustrated in equation (2). In general,

the features and techniques applied in weblogs were used in our system as guides only and to know which factors affect the credibility of this genre of web information systems. Moreover, while applying these measures to Twitter, they were tailored to work adequately with the limitation and specific features of Twitter.

Finally, as this work is considered a first attempt to measure Twitter credibility, we think that the results are acceptable and the system could be applied to any kind of news. Yet, there are a number of limitations for this work which can be summarized as follows:

(1) The credibility formula is experimental and it is subject to further investigation.

(2) The similarity thresholds values were based on (Juffinger *et al.*, 2009), thus further experiments are suggested to derive these values specifically for Twitter.

(3) The collection of credible news web sites or IW, were done manually. Therefore, we suggested automating this process by using algorithms such as the one experimented by (Nagura *et al.*, 2006) to rate the credibility of news articles on the Web or using machine learning techniques for IW extraction.

Notes

1. Khojah stemmer. Retrieved August 2010, available at: http://zeus.cs.pacificu.edu/shereen/research.htm#stemming
2. Stanford POS tagger. Retrieved August 2010, available at: http://nlp.stanford.edu/software/tagger.shtml

References

Akamine, S., Kato, Y., Inui, K. and Kurohashi, S. (2008), "Using appearance information for web information credibility analysis", *International Symposium on Universal Communication*, IEEE Computer Society, Los Alamitos, CA, pp. 363-5.

Conrad, J.G., Leidner, J.L. and Schilder, F. (2008), "Professional credibility: authority on the web", *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, ACM, Napa Valley, CA, pp. 85-8, available at: http://portal.acm.org/citation.cfm?id=1458548 (accessed August 12, 2010).

Fogg, B.J. (2003), "Prominence-interpretation theory: explaining how people assess credibility online", *CHI '03 Extended Abstracts on Human Factors in Computing Systems*, ACM, Ft. Lauderdale, FL, pp. 722-3, available at: http://portal.acm.org/citation.cfm?id=765951 (accessed August 12, 2010).

Fogg, B.J., Marshall, J., Laraki, O., Osipovich, A., Varma, C., Fang, N., Paul, J., Rangnekar, A., Shon, J., Swani, P. and Treinen, M. (2001), "What makes web sites credible? A report on a large quantitative study", *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM, Seattle, WA, pp. 61-8, available at: http://portal.acm.org/citation.cfm?id=365037 (accessed August 12, 2010).

Grinev, M., Grineva, M., Boldakov, A., Novak, L., Syssoev, A. and Lizorkin, D. (2009), "Sifting micro-blogging stream for events of user interest", *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, ACM, Boston, MA, p. 837, available at: http://portal.acm.org/citation.cfm?id=1572157 (accessed August 12, 2010).

Harrag, F., Hamdi-Cherif, A., Al-Salman, A. and El-Qawasmeh, E. (2009), "Experiments in improvement of Arabic information retrieval", paper presented at the 3rd IEEE International Conference on Arabic Language Processing (CITALA'09), Rabat, Morocco.

Hilligoss, B. and Rieh, S.Y. (2008), "Developing a unifying framework of credibility assessment: construct, heuristics, and interaction in context", *Inf. Process. Manage.*, Vol. 44 No. 4, pp. 1467-84.

Honey, C. and Herring, S. (2009), "Beyond microblogging: conversation and collaboration via Twitter", *The 42nd Hawaii International Conference on System Sciences*, IEEE Computer Society, Big Island, HI, pp. 1-10.

House, N. (2004), "Weblogs: credibility and collaboration in an online world", available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.126.6034 (accessed August 12, 2010).

Iding, M.K., Crosby, M.E., Auernheimer, B. and Klemm, E.B. (2008), "Web site credibility: why do people believe what they believe?", *Instructional Science*, Vol. 37 No. 1, pp. 43-63.

Java, A., Song, X., Finin, T. and Tseng, B. (2007), "Why we Twitter: understanding microblogging usage and communities", *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, Vol. 12, ACM, San Jose, CA, pp. 56-65, available at: http://portal.acm.org/citation.cfm?id=1348556 (accessed August 12, 2010).

Juffinger, A., Granitzer, M. and Lex, E. (2010), "Blog credibility ranking by exploiting verified content", *Proceedings of the 3rd Workshop on Information Credibility on the Web*, Vol. 12, ACM, Madrid, pp. 51-8, available at: http://portal.acm.org/citation.cfm?id=1527005 (accessed August 12, 2010).

Kaczmarek, A. (2008), "Automatic evaluation of information credibility in Semantic Web and Knowledge Grid", paper presented at the 4th International Conference on Web Information Systems and Technologies (WEBIST). Madeira, Portugal.

Kamthan, P. (2008), "A framework for the active credibility engineering of web applications", *International Journal of Information Technology and Web Engineering* (*IJITWE*), Vol. 3 No. 3, pp. 17-27.

Kidawara, Y. (2008), "Information credibility analysis of web content", *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, ACM, Napa Valley, CA, pp. 3-4, available at: http://portal.acm.org/citation.cfm?id=1458530 (accessed August 12, 2010).

Krishnamurthy, B., Gill, P. and Arlitt, M. (2008), *Proceedings of the first Workshop on Online Social Networks*, ACM, Seattle, WA, pp. 19-24, available at: http://portal.acm.org/citation.cfm?id=1397735.1397741 (accessed August 12, 2010).

Kwon, K., Cho, J. and Park, Y. (2009), "Multidimensional credibility model for neighbor selection in collaborative recommendation", *Expert Systems with Applications*, Vol. 36 Nos 3, pp. 7114-22 (Part 2).

Lazar, J., Meiselwitz, G. and Feng, J. (2007), *Understanding Web Credibility: A Synthesis of the Research Literature*, Now Publishers, Delft.

Lopes, R. and Carriço, L. (2008), "On the credibility of Wikipedia: an accessibility perspective", *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, ACM, Napa Valley, CA, pp. 27-34, available at: http://portal.acm.org/citation.cfm?id=1458536 (accessed August 12, 2010).

Mishne, G.A. (2007), *Apply Text Analysis for Blogs*, University of Amsterdam, Amsterdam.

Parikh, R. and Movassate, M. (2009), "Sentiment analysis of user-generated twitter updates using various classification techniques", *Business Marketing*, June.

Nagura, R., Seki, Y., Kando, N. and Aono, M. (2006), "A method of rating the credibility of news documents on the web", *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Seattle, WA, ACM, pp. 683-4, available at: http://portal.acm.org/citation.cfm?id=1148316 (accessed August 12, 2010).

Nakamura, S., Shimizu, M. and Tanaka, K. (2008), "Can social annotation support users in evaluating the trustworthiness of video clips?", *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, ACM, Napa Valley, CA, pp. 59-62, available at: http://portal.acm.org/citation.cfm?id=1458542 (accessed August 12, 2010).

Narayanan, A. and Shmatikov, V. (2010), "De-anonymizing Social Networks", *Proceedings of the 2009 30th IEEE Symposium on Security and Privacy*, IEEE Computer Society, Oakland, CA, pp. 173-87, available at: http://portal.acm.org/citation.cfm?id=1608132 (accessed August 12, 2010).

Phelan, O., McCarthy, K. and Smyth, B. (2009), "Using Twitter to recommend real-time topical news", *Proceedings of the Third ACM Conference on Recommended Systems*, Vol. 12, ACM, New York, NY, pp. 385-8, available at: http://portal.acm.org/citation.cfm?id=1639714.1639794 (accessed August 12, 2010).

Rieh, S.Y. and Danielson, D.R. (2007), "Credibility: a multidisciplinary framework", *Annual Review of Information Science and Technology*, Vol. 41 No. 1, pp. 307-64.

Rubin, V. and Liddy, E. (2006), "Assessing credibility of weblogs", *AAAI Symposium on Computational Approaches to Analysing Weblogs* (*AAAI-CAAW*), pp. 187-90, available at: http://www.aaai.org/Papers/Symposia/Spring/2006/SS-06-03/SS06-03-038.pdf

Sankaranarayanan, J., Samet, H., Teitler, B.E., Lieberman, M.D. and Sperling, J. (2009), "TwitterStand: news in tweets", *Proceedings of the 17th ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems*, ACM, Seattle, WA, pp. 42-51, available at: http://portal.acm.org/citation.cfm?id=1653781 (accessed August 12, 2010).

Tsagias, M., Larson, M., Weerkamp, W. and De Rijke, M. (2008), "PodCred: a framework for analyzing podcast preference", *Proceedings of the 2nd ACM workshop on Information credibility on the web*, ACM, Napa Valley, CA, pp. 67-74, available at: http://portal.acm.org/citation.cfm?id=1458545 (accessed August 12, 2010).

Tseng, S. and Fogg, B.J. (1999), "Credibility and computing technology", *Communications of the ACM*, Vol. 42 No. 5, pp. 39-44.

Ulicny, B. and Baclawski, K. (2007), "New metrics for newsblog credibility", *Proceedings of the First International Conference on Weblogs and Social Media* (*ICWSM07*), available at: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.157.6230 (accessed August 12, 2010).

Wanas, N., El-saban, M., Ashour, H. and Ammar, W. (2008), "Automatic scoring of online discussion posts", *Proceedings of the 2nd ACM Workshop on Information Credibility on the Web*, ACM, Napa Valley, CA, pp. 19-26, available at: http://portal.acm.org/citation.cfm?id=1458534 (accessed August 12, 2010).

Weerkamp, W. and De Rijke, M. (2008), "Credibility improves topical blog post retrieval", *Annual Meeting-Association For Computational Linguistics*, pp. 923--931.

Weimer, M., Gurevych, I. and Mühlhäuser, M. (2007), "Automatically assessing the post quality in online discussions on software", *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, Association for Computational Linguistics, Prague, pp. 125-8, available at: http://portal.acm.org/citation.cfm?id=1557806 (accessed August 12, 2010).

About the authors

Hend S. Al-Khalifa is Assistant Professor at Information Technology Department, College of Computer and Information Sciences, King Saud University, Riyadh, Saudi Arabia. She received her PhD degree in Computer Science (2007) from Southampton University, UK. Her areas of interest include web technologies (i.e. Semantic Web/Web 2.0), technology enhanced learning (e.g. e-learning, adaptive hypermedia, etc), computer for people with special needs and Arabic language and computers. Hend S. Al-Khalifa is the corresponding author and can be contacted at: hendk@ksu.edu.sa

Rasha M. Al-Eidan received her MSc Degree in Computer Science from King Saud University in 2010. Her area of interest includes web technologies and e-learning.