A conceptual model of the automated credibility assessment of the volunteered geographic information

N H Idris[1,4], M J Jackson[2]and M H I Ishak[3]

[1] Sustainability Research Alliance and Department of Geoinformation, Universiti Teknologi Malaysia (UTM), 81310 UTM, Johor, Malaysia
[2]Nottingham Geospatial Institute, Nottingham Geospatial Building Triumph Rd, The University of Nottingham, Nottingham NG7 2TU,United Kingdom
[3]Infocom Research Alliance, Universiti Teknologi Malaysia (UTM), 81310 UTM, Johor, Malaysia

Email: hawani@utm.my

Abstract. The use of Volunteered Geographic Information (VGI) in collecting, sharing and disseminating geospatially referenced information on the Web is increasingly common. The potentials of this localized and collective information have been seen to complement the maintenance process of authoritative mapping data sources and in realizing the development of Digital Earth. The main barrier to the use of this data in supporting this bottom up approach is the credibility (trust), completeness, accuracy, and quality of both the data input and outputs generated. The only feasible approach to assess these data is by relying on an automated process. This paper describes a conceptual model of indicators (parameters) and practical approaches to automated assess the credibility of information contributed through the VGI including map mashups, Geo Web and crowd - sourced based applications. There are two main components proposed to be assessed in the conceptual model – metadata and data. The metadata component comprises the indicator of the hosting (websites) and the sources of data / information. The data component comprises the indicators to assess absolute and relative data positioning, attribute, thematic, temporal and geometric correctness and consistency. This paper suggests approaches to assess the components. To assess the metadata component, automated text categorization using supervised machine learning is proposed. To assess the correctness and consistency in the data component, we suggest a matching validation approach using the current emerging technologies from Linked Data infrastructures and using third party reviews validation. This study contributes to the research domain that focuses on the credibility, trust and quality issues of data contributed by web citizen providers.

1. Introduction

The current Web 2.0 revolution and the Digital Earth vision have made big impact on the culture of mapping. This new 'geo' landscape incorporates aspects described by a number of new terms and concepts, including neogeography [1] and volunteered geographic information (VGI) [2]. Under this revolution, professional geoliterate users are not the only group that is active in mapping activities. Users now not only use authorized data sources but also from VGI sources. This new type of data source is becoming more practical due to its accessibility and locality coverage. Although there is a trade-off in terms of quality, users tend to use and 'believe' this information.

This situation is in contrast to conventional mapping where the data supplied by authoritative data source. The data usually comes with standard metadata that has been produced for users to assess the fitness of the data for their purposes [3]. This is less likely to happen in the case of data and

[4] To whom any correspondence should be addressed.

information presented in neogeography based tools. For example in map mashup context, the data are mashed up from various sources, where the metadata are recorded in informal and unstructured formats.

Previous studies have identified the low influence of metadata [4] and a high influence of credibility (trust) labelling when users judging the credibility of map mashup information [5]. These findings were supported by a few studies from other domains, for example [6,7] that have identified the high influence of visual design when users judge the credibility of online information. Other collaborative crowd source based user generated contents applications such OpenStreetMap has its own moderator (gatekeeper) to deal with the issues of vandalism, copyright violation and disputes. Such application has mechanisms to validate and correct the errors by using the ability of the crowd to converge on the truth [8]. There is no gatekeeper, however to control the correctness of information presented on other VGI medium such as map mashup and Geoweb applications.

Research in GIS domain has actively proposed the use of graphic visualisation to increase users' awareness of the quality and uncertainty of the data they use. A study by Devillers et al. [9] have proposed a tool that uses a colour coded traffic light scheme (CCTL) label to present the accuracy of the features, layers and datasets in a desktop, web and mobile GIS application. Wilson and Graham [10] argue that it is crucial to provide information on uncertainties in data when disseminated thorough a neogeography tool to avoid misleading interpretation by citizen users.

The presence of an automated credibility tool which could check and assess the level of correctness of VGI can assist the decision processes during a post-disaster relief as well as protecting users from propaganda, incorrect, misleading and invalid sources of information. This paper discusses the proposed conceptual model of the parameters to evaluate the credibility (believability) of the VGI and neogeography based medium. Section 2 presents a comprehensive review of elements related to credibility that have been identified in several domains where in the end of this section, credibility elements specifically for VGI are proposed. Section 3 then discuss the practical approaches that recommended assessing these credibility elements in automated manner.

2. Credibility elements

Credibility is synonymous with believability [11,12]. Credibility influences the viewer's perception of believability in the information conveyed by the object. The object of assessment may refer to the source, message, or the media itself. In the literature, these credibility cues have been examined in various domains, including information science (IS), human computer interaction (HCI) and health information science although slightly different semantic terms are used.

Table 1 below summarizes credibility related elements drawn from a few studies in different domains. These elements then become a basis to the selection of practical credibility elements to assess VGI in automated manner.

Table1. Summary of credibility elements identified from a few studies from different domain.

Studies	Domains	Terms	Elements related to Credibility
[13]	Information Science(IS)	Website credibility	Information focus, name recognition/affiliation, links, commercial interest, reference, information design, currency, design look, expertise, bias, inaccuracy, information clarity, tone, corroboration
[6]	Human Computer Interaction (HCI)	Website credibility	Design look, Information Design/structure, Information focus, Motive, Usefulness, Accuracy reputation, Advertising, Bias, Writing tone. Identity of site sponsor Functionality, customer service, information clarity, readability, affiliation

Table1. continued

Studies	Domains	Terms	Elements related to Credibility
[14]	Geospatial	Trust	Proximity to author's location to the event, Reputation (individual personality)

[15]	Health Information Science	Indicator of accuracy	Copyright acknowledged, exclamation points, citation, high number of in-links,Website domain, up-to-date, HON-code logo, advertising, disclose author, disclose contact information of errors in spelling

2.1. The practical credibility elements

The main practical indicators discussed below are elements of credibility recommended for assessing VGI. These indicators have been drawn from the literature and recommended to be assessed in an automated manner. The main indicators proposed in the credibility rating model comprise two main components:

1) Metadata component – In this component, the elements to be evaluated are related to the source and hosting (website). Figure 1 presents the proposed elements in this component.

2) Data component – In this components, the proposed elements are drawn from the third party reviews either from independent and/or dependent sources and the consistency and correctness related to spatial data. Figure 2 presents the proposed elements in data component.

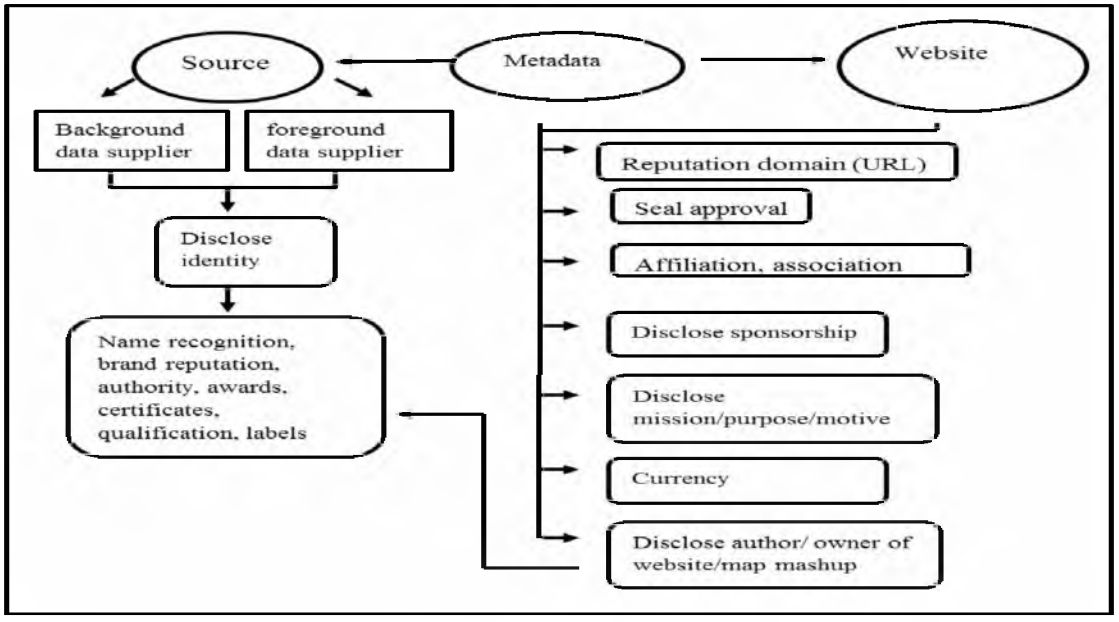

Figure 1. The proposed elements to be assessed in metadata component.

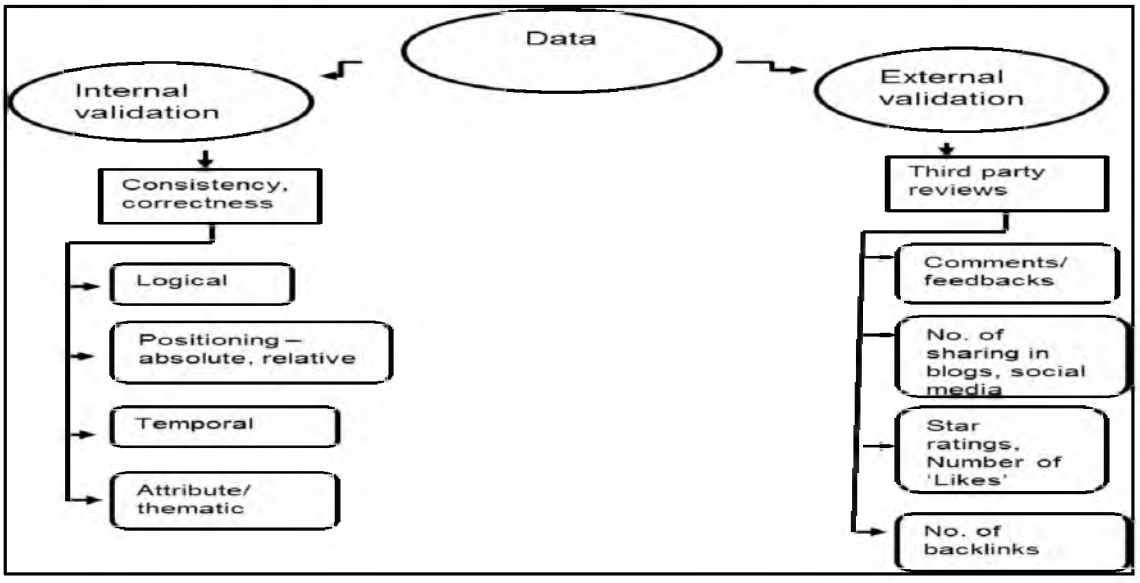

Figure 2. The proposed elements to be assessed in data component.

2.2. The practical approaches to assess metadata component

There are several practical ways to evaluate the elements suggested in the metadata component. One way is to detect the metadata indicator using rule based concept. For example, Wang and Richard [16] have proposed a rule based line classifications to enable the recording and detecting of metadata indicators for online health information. In that study, a set of rules for recording quality-related metadata by web developers is proposed.

Another way is to develop a digital metadata vocabulary by the application of machine based learning. In this approach, metadata is developed through a supervised learning method, similar in technique to the automatic recognition and filtering of spam email applications. This method is more effective than rule-based line classification. A good corpus of training data can be generated because of its ability to capture and model events that individual experts may not have identified; these events can then be evaluated in coherent ways [17]. In this method, a huge collection of metadata to describe data/information in mapping applications, including VGI, has to be compiled as a training dataset. This dataset will be used to develop a corpus, with patterns of texts or events corresponding to specific categories of metadata indicators available for examination.

2.3. The practical approaches to assess data component

There are several practical ways to assess the elements suggested in the data component in automated manner. Although there is still a long way to go, a few challenges have been identified. First is to find appropriate or matching datasets to compare and evaluate data/information. Data matching has been of research interests in geospatial domain. For example a study by Koukoletsos et al. [18] has demonstrated the ability to match feature (linear) based data using automated method. That study proposed automated matching approach by considering the nature of VGI data, particularly in terms of the heterogeneity of data and the constraints of attributes.

Another approach to find the matching references (dataset) is to rely on a currently emerging development in the realm of the Semantic Web, namely linked data infrastructure. Linked Data Web is a technology to interlink data, information, concepts and facts on the WWW. The Possible matching data could be drawn from governmental and commercial agencies that publish some of their data through the Linked Data Web infrastructure. It has been argued that current developments in the Semantic Web could not fully support interlinked data from government and enterprise datasets due to a lack of research into the best practical approaches to support large scale adoption. Nonetheless, a few attempts have been made by major organizations, such as BBC media and the UK Government, including the Ordnance Survey of Great Britain (OSGB), to publish some of their data through the Linked Data Web infrastructure. Possible matching data could also draw from transforming VGI efforts such as OpenStreetMap dataset into an RDF data model to interlink

with Linked Data infrastructure. Standler et al. and Auer et al. [19,20] for example demonstrates interlink OSM dataset into DBpedia and Geonames linked data infrastructure. There is a promise in the near future, for other geospatial datasets and information to be published in the RDF data model and interlinked in Linked Data infrastructure, particularly among government agencies that have supported and will support open data policies.

Comparison from independent reports can be used to verify the correctness of data or information. For example during the Hurricane Sandy, there a lot of sources disseminate the related news, including online news, governmental websites, search and rescue disaster unit websites, individual blog, social bookmarking sites (Delicious.com, Digg.com) and social networking sites (Facebook and Twitter). These online sources could be used to verify the information. Independent reports such as from 'booking.com', 'tripadvisor.com' may be relevant to validate information related to places of interest. As argue by Metzger and Flanagin [21] combining recommendation and opinions from users generated contents could become 'evidence' to support facts, data or information but bases on individual experiences and local knowledge. Machine learning algorithms could be used to detect relevant keywords or area coverage in external sites before validation of data or information can be conducted.

3. Conclusion and future work

In conclusion, there is a need of an evaluation tool to assess the level of believability (credibility), correctness and consistency of data and information presented through VGI medium. This paper proposed credibility elements of metadata and data components that can be used to assess this form of data. The recommended approaches to assess the elements in automated manner include machine based learning, third party validation and linked data infrastructure. These approaches could bring together data/information between and from different sources, and immensely useful in facilitating the automated process of searching for and matching data/information for accuracy and correctness validation. The next stage of this study is to develop the automated tools and test the proposed conceptual model. The implication of conceptual model proposed in this study could be applied to assess data and information particularly when metadata are presented in informal and unstructured format or the sources of data are invalidated.

Acknowledgement

This project was funded by the Universiti Teknologi Malaysia (UTM) and Ministry of Higher Education Malaysia (MOHE) under Research University Grant vote no. 08J94.

References

[1] Turner A J 2006 Introduction to Neoegeography. O'Reilly Shortcut [Online] Available:http://pcmlp.socleg.ox.ac.uk/sites/pcmlp.socleg.ox.ac.uk/files/Introduction_to_Neogeography.pdf [AccessedApril2012]

[2] Goodchild M F 2007 Citizens as Sensors: The World of Volunteered Geography *GeoJournal* **69** 211-221

[3] Elwood S, Goodchild M F and Sui D Z 2012 Researching Volunteered Geographic Information: Spatial Data, Geographic Research, and New Social Practices *Annals of the Association of American Geographers* **102** 571-590

[4] Idris N H, Jackon M J and Abrahart, R J 2011 Map Mashups: What looks good must be good? GIS Research UK Conference (GISRUK), Portsmouth UK, 27-29

[5] Idris N H, Jackon M J and Abrahart, R J 2011 Colour Coded Traffic Light Labelling: An Approach to Assist Users in Judging Data Credibility in Map Mashup Applications. In Cidalia C. Fonte, Luisa Goncalves & Gil Goncalves (Eds), *Proc.of the 7th International Symposiumon Spatial Data Quality* Coimbra Portugal: INESC Coimbra 201-206

[6] Fogg B J ,Cathy S, David R D, Leslie M, Julianne Sand Ellen R T 2003 How do users evaluate the credibility of Websites?: a study with over 2,500 participants. *Proc of the Conference on Designing for User Experiences*. San Francisco, California: ACM.

[7] Albert W A and Van D G T M 2011 Color Matters: Color as Trust worthiness Cuesin WebSites. *Technical Communication* **58** 149-160

[8] Goodchild M F and Li L 2012 Assuring the quality of volunteered geographic information.

Spatial Statistics **1** 110-120
[9] Devillers R, Bedard Y, Jeansoulin R and Moulin B 2007.Towards Spatial Data Quality Information Analysis Tools for Expert Assessing the Fitness for Use of Spatial Data. *Int. J. of Geographic Information Science* **21** 261-282
[10] Wilson M Wand Graham M 2013 Neogeography and volunteered geographic information: a conversation with Michael Good child and Andrew Turner. *Environmentand Planning* A **45** 10-18
[11] Fogg B Jand T seng H 1999 The elements of computer credibility. *Proc. Of the SIGCHI conferenceon Human factors in computing systems* Pittsburgh, Pennsylvania, UnitedStates: ACM.
[12] Flanagin A J and Metzger M J , 2008 The credibility of Voluntereed Geographic Information. *GeoJournal* **72** 137-148
[13] Iding M K, Crosby M E, Auernheimer B and Klemm E B 2009 Website Credibility: Why Do People Believe What They Believe. *Instructional Science* **37** 43-63
[14] Bishr Mand Mantelas L 2008 A trust and reputation model for filtering and classifying knowledge about urban growth *Geo Journal* **72** 229-237
[15] Fallis D and Fricke M 2002 Indicators of accuracy of consumer health information on the Internet: a study of indicators relating to information for managing fever in children in the home. *J. of the American Medical Informatics Association* **9** 73-79
[16] Wang Yand Richard R 2007 Rule-based Automatic Criteria Detection for Assessing Quality of Online Health Information *Int. Conf Addressing Information Technology and Communication in Health (ITCH)* Victoria Canada
[17] Gaudinat A, Grabar N, Boyer C, Bellazzi R, Abu-Hanna A and Hunter J 2007 Automatic Retrieval of Web Pages with Standards of Ethics and Trust worthiness Withina Medical Portal: What a Page Name Tells Us Lecture Notes in Computer Science **4594** 185-189
[18] Koukoletsos T, Haklay Mand Ellul C 2012 Assessing Data Completeness of V G I through an Automated Matching Procedure for Linear Data *Transaction in GIS* **164** 477-498
[19] Auer S, Lehmann J. and Hellmann S 2009 Linked Geo Data: Adding a Spatial Dimension to the Web of Data *Lecture Notes in Computer Science* **5823** 731-746
[20] Stadler C, Lehmann J, Höffner K,and Auer, S 2012 Linked Geo Data: A Core for a Web of Spatial Open Data. *Semantic Web* **3** 333-354
[21] Metzger M J, and Flanagin A J 2011 Using Web 2.0 Technologies to Enhance Evidence-Based Medical Information *J. of HealthCommunication* **16** 45-58