Semi-Automated Metadata Detection for Assessing the Credibility of Map Mashup

Nurul H IDRIS, Mohamad N SAID, Mohamad HI ISHAK, Mohamad G HASHIM, Zamri ISMAIL, Malaysia
Mike J JACKSON, United Kingdom

Key words: Map Mashup, Web Crawler, Metadata, Credibility, Trust

SUMMARY

Current Web 2.0 technologies enable the easy sharing of geospatial data from multiple on-line sources. With free mapping APIs, online maps can be created without the need for high cost software, meaning that a map can be created by both the professional as well as the web-enabled citizen. In a traditional approach, metadata is used to assess the fitness of the data to meet the purpose of usage. In this new mapping landscape, metadata can be used to evaluate the credibility (believability) of information and to generate trust in the sources. However, in this new domain, the embedded metadata is typically partial, informal and unstructured. This paper demonstrates the use of a web crawler to discover supplementary metadata. This semi-automated detection is one of the components needed to support a framework for assessing the credibility of information presented in map mashups, particularly those presenting crowd-sourced data. This framework may then be used to tackle the issues of credibility and trust related to Web 2.0 mapping applications.

Semi-Automated Metadata Detection for Assessing the Credibility of Map Mashup

Nurul H IDRIS, Mohamad N SAID, Mohamad HI ISHAK, Mohamad G HASHIM, Zamri ISMAIL, Malaysia
Mike J JACKSON, United Kingdom

1. INTRODUCTION

The Web has been a medium for the dissemination of information since the time of its invention. The emergence of map mashup technology is one of the outcomes driven by the Digital Earth vision and the release of the Google Maps API in 2005. The term mashup was originally used to describe the blending of musical tracks to create new forms of song; the term now also refers to websites that weave data from different sources into new integrated user services (Batty et al., 2010:2). At the time of writing, more than 2000 map mashup applications are identified by the Programmableweb (2013) website portal.

Under this revolution, a web citizen typically not only uses authorized data sources but also draws from volunteered geographic information (VGI) and other crowd-sourcing sources when seeking geospatial based information. These new data sources are becoming more practical due to their accessibility and locally detailed coverage. Although there may be 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 is supplied by an 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 (Elwood et al., 2012). This is less likely to happen in the case of data and information presented in a map mashup context where the data are mashed up from various sources and recorded in informal and unstructured formats. In this new mapping domain, metadata is used to evaluate the credibility (believability) of information and to check whether the sources can be trusted.

Research related to credibility has become of interest in several domains, including communication, information science, marketing, psychology, interdisciplinary efforts in human computer interaction (HCI) and, currently, Web 2.0 applications. Issues of credibility in user-generated spatial content, such as OpenStreetMap and free web mapping service applications in Google Earth, have also been raised by several authors, including Flanagin and Metzger (2008) and Goodchild (2008).

Credibility is tightly coupled with believability (Flanagin and Metzger, 2008; Fogg and Tseng, 1999). Credibility is an intangible concept and is related to a user's perception of an object of assessment. 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. The main primary dimensions of credibility discussed in the literature are trustworthiness and expertise (Fogg and Tseng, 1999). As described by Fogg and Tseng (1999:80), 'trustworthiness is defined by the terms well-intentioned, truthful, and

unbiased; the trustworthiness dimension of credibility captures the perceived goodness or morality of the source. Expertise however is defined by the terms such as knowledgeable, experienced, and competent; this dimension captures the perceived knowledge and skill of the source'. Previous studies have identified the low influence of metadata (Idris et al., 2011a) and a high influence of credibility (trust) labelling when users judge the credibility of map mashup information (Idris et al., 2011b). These findings were supported by a few studies from other domains, for example see Fogg et al. (2003) and Albert and Van (2011) that have identified the high influence of visual design when users judge the credibility of online information. Other collaborative crowd-source based user generated content applications, such as OpenStreetMap, have their own moderators (gatekeepers) to deal with the issues of miss and disinformation, copyright violation and disputes. Such applications have mechanisms to help validate data and correct errors by using the characteristic of the crowd to converge on the truth (Goodchild and Li, 2012). There is typically no gatekeeper, however, to control the correctness of information presented on map mashup applications.

Idris et al. (2013a) proposed elements of metadata to measure and assess credibility of neogeography based and volunteered geographic information (VGI) applications. Idris et al. (2013b) also proposed a framework that consisted of three major components – a web crawler, a digital metadata vocabulary and a credibility index - to assess credibility of map mashups through metadata assessment. The purpose of this paper is to demonstrate the first component, a web crawler, in terms of its ability to detect metadata parameters that may assist in the indexing and rating of the map mash-up by the subsequent components. Section 2 highlights the procedures of the web crawler to detect the metadata. Section 3 presents the results of the detection by the web crawler. Section 4 summaries this paper, discusses the limitations and suggests future research directions.

2. METHODS

2.1 Web data harvesting

Web data harvesting, web scraping or web extraction are similar terms to describe an approach in the domain of Web data mining to discover useful data and knowledge from web pages. These techniques have emerged due to the high potential for extracting valuable data and information from the web. The way information and data are presented on the web in a human readable format limits the mechanism of a computer to automate the data and information extraction. By the use of a web crawler, data and information on a website can be extracted in an automated manner.

A web crawler is commonly used in an information retrieval domain such as a search engine application to find similar matches of web pages according to keywords input by users. Other terms which may be used in this context are spiders, worms, robots, walkers and wanderers. For example, the Google engine is reported to use multiple distributed web crawlers in its large scale search engine to support the query search and to filter the results according to relevancy criteria (Brin and Page, 2012). The infrastructure consists of major components, including web crawler, URL server, indexer, repository, barrels, lexicon database, anchor and sorter. A URL server sends a list of URLs to be fetched to the crawlers

and the fetched web pages are then sent to the store server. The stores server then compresses and stores the web pages in a repository before an indexer processes the documents and stores the important information about them in an anchor file. This information is used to compute the PageRank, which is an algorithm mechanism, by Google to prioritise the relevancy of the result search (see Carr, 2006).

In literature, a few studies have proposed a framework to detect metadata indicators. One related study that detects structured metadata indicators is Wang and Richard (2007). That study proposed a rule-based line classification to enable the recording and detecting of metadata indicators for online health information. In this study, a set of rules for recording quality-related metadata by web developers was proposed. For example, a set of rules regarding which indicator value to scan, which section/block on the document and which patterns to be scanned need to be predetermined before automated assessment by web crawler can be conducted. This rule-based approach is requires rules that need to be pre-determined by experts to be effective. If the rules are less comprehensively defined, there may be certain events that cannot be detected by web crawlers. A study by Olfat et al. (2012) also relied on a rule based approach to record updated metadata through GML Application Schema and OGC web services.

There are several techniques that could be used by a web crawler to extract data and information from a website. The concept of a web crawler is slightly similar to the concept of making a query to a database. It uses advance query types that enable a search of the surface and the back codes that are used to display the data and information. The types of search that could be used by a web crawler including exact match, wildcard match, regular expressions, Xpath and scrape. Xpath is a XML path language to query nodes from a XML document. Scrape is a more advanced query type; this mode provides an algorithm to generate a query and extract specific information from a website; this technique implements a supervised machine learning approach where a crawler will learn to detect code expression patterns from the training patterns that have to be set up before a crawler conducts a search (Inspyder, 2013).

This study demonstrates the ability of a web crawler to detect metadata in order to assess the credibility of map mashups through the use of off-the-shelf commercial software, namely Mozenda™ and Inspyder Crawler™. The first tool, Mozenda™, was used in this study due to its ability to instruct the supervised machine learning of a web crawler to detect metadata through screen scraping efficiently. The latter tool was used to demonstrate other mechanisms to instruct the web crawler, for example by searching keywords of metadata as well as by pre-defined code expressions.

2.2 Defining measurable indicators

Previous work (Idris et al., 2013b) has identified several types of metadata that could be used to assess the credibility of map mashup which are adapted from the metadata standard ISO 19115, FGDC, the Dublin Core as well as credibility indicators from various other domains. The proposed parameters include disclosing author/creator of the map mashup, currency – publication date, mission, purpose or motivation, currency for the period under consideration, currency – last updated date, maintenance and update frequency statement,

contact information, source, disclosure of identity of background and foreground data suppliers, domain URL reputation, data supplier reputation, seals of approval, affiliation/association and sponsorship.

In a previous study, Wang and Richard (2007) proposed rule-based automatic criteria detection in which the name, value and location of each indicator have to be pre-defined. For a web crawler that implements supervised machine learning, the location of each criterion is not required to be pre-defined; only the name (i.e. metadata indicators) and its value are defined. The format of code expressions used to store the metadata, which is either structured or unstructured, influences the procedures of a web crawler to detect indicators. Table 1 presents the common code expressions and possible values for the tested parameters. For structured metadata, only the names of indicators are pre-defined. Meanwhile, for unstructured metadata the name and the value of indicators need to be pre-defined; a training dataset for a list of possible values of the indicator is required to assist in this approach.

2.3 Detecting indicators

The processes of detecting metadata indicators using a web crawler in this study are illustrated in Figure 1 (light grey coloured). There are two main steps in detecting the metadata indicators, which are: 1) identifying and defining the metadata parameters; 2) capturing the parameters. Dark grey coloured shapes in Figure 1 present the processes that will be activated after an indicator is detected.

Table 1 Metadata indicators and their common code expressions and possible values

Name (metadata indicator)	Common code expressions	Possible values (keywords)
Currency – publication date	structured	Date, published data
Currency – last updated date	structured	Date, last updated date
Currency for the period under consideration	unstructured	Timeline, temporal, period
Disclose author/creator map mashup	structured	Published by, created by, developed by, author,
Contact information	structured/ unstructured	About us, email, organization, telephone (phone), fax number,
Disclose identity of background data	unstructured	Copyright
Disclose identity of foreground data supplier	structured	Source, supplied by, origin,
Reputation domain URL	structured	URL http://
Affiliation/association	structured	Hyperlinks
Seal approval	unstructured	Approved by,
Sponsorship	unstructured	Sponsored by, funded by,
Mission/purposes/ motive	unstructured	About us, Objectives, purposes, goal

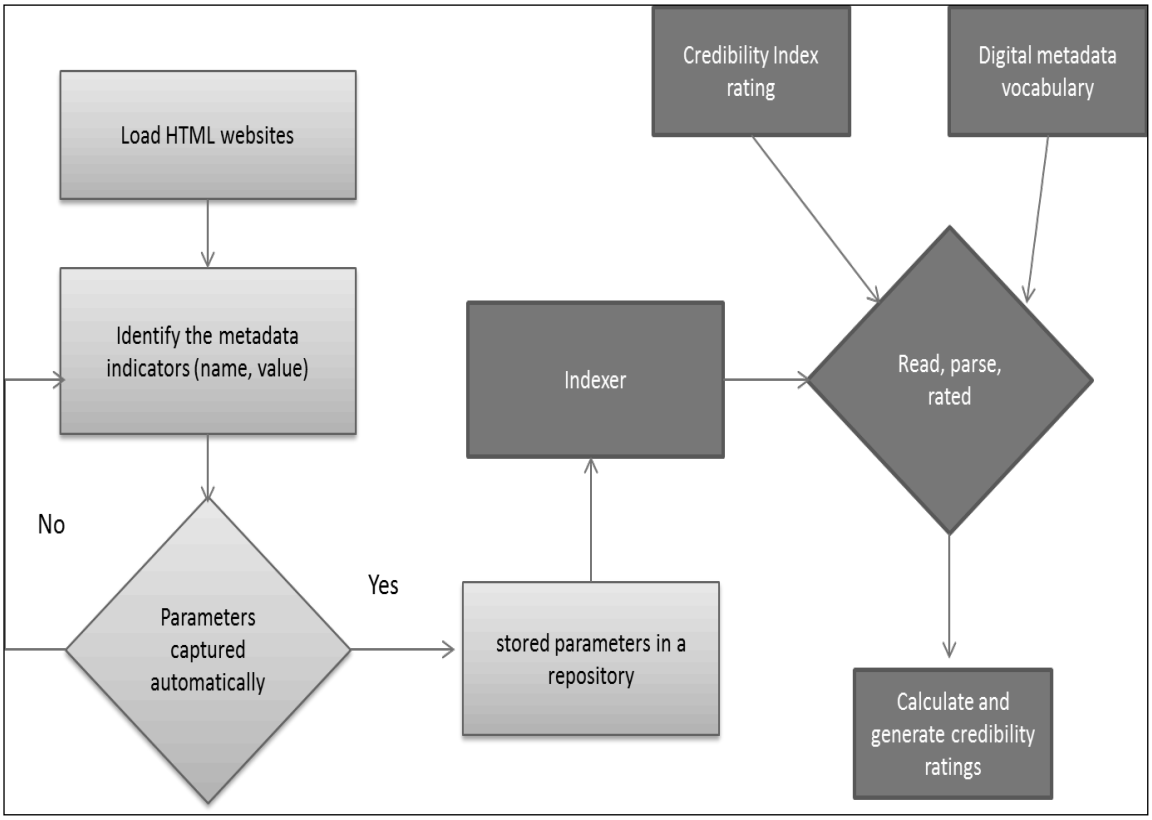

Figure 1: The processes of detecting metadata indicators

2.3.1 Identify and define the parameter

Identifying and defining a metadata parameter on map mashup applications requires preprocessing steps for detection. This preprocessing is essential to produce a pre-defined parameter of name (item) and value for a measurable metadata indicator and the possible pattern of code expressions for the web crawler to search. The ideal procedure to conduct this preprocessing phase is to have a defined digital metadata vocabulary and a training dataset that could be used as a basis to produce these pre-define parameters and patterns. However, for this pilot study, such a digital metadata vocabulary and training dataset had not yet been developed. The basis of pre-defined parameters and patterns used in this study was therefore based on the literature and observations of 10 samples of map mashup applications.

After the HTML web pages are loaded, human assistance is required to define the parameters that need to be detected by a web crawler. For example, the author or creator of the map, date of publication, objectives of the map mashup application, publisher, the source(s) of base map and foreground data and the URL domain of a website. The indicators and patterns that express the parameters will be identified and stored in a database.

2.3.2 Capturing the parameters

Once the patterns of metadata indicators have been pre-defined, a crawler will search the metadata indicators at the surface of the web pages and in non-hidden back codes in the form of either HTML, CSS, Xpath or javascript that is used by web documents to present the metadata. It will crawl the documents, including the web pages and embedded maps, to capture the parameter values that are matched with those pre-defined at the preprocessing step. For metadata that is commonly displayed in a structured format (see Table 2), screen scraping technique could be used to intelligently detect the indicators' values through the sample of training patterns that have been previously identified. In this technique, only metadata indicators are defined; the value of the parameter does not need to be pre-defined. The crawler will intelligently capture all the values that are associated with the indicators. Moreover, the Mozenda™ tool has a collection of pre-defined regular expressions that can intelligently detect the regular expressions for common structured indicators, including URL website domain, date and emails. On the other hand, for metadata that is commonly presented in unstructured format, the value of parameters first need to be identified. Following this a direct detection technique, for example by using exact match query type, could be applied to find the matching parameter values.

After the crawler has found the possible matched parameters, it will store the matching values in a database repository. These values will be used by the indexer component. The Indexer will read and parse the values by referring to the digital metadata vocabulary component. This function of digital metadata vocabulary is similar with other lexicon database. This component will provide the basis to match the ontology and semantic similarities of the values detected by the crawler. After a matching value is found in the vocabulary database, the value will then be rated and calculated by the credibility index rating component. This rating index component consists of the consensus expert views of the ratings for each parameter. Calculation of the overall ratings will total up the rating for each parameter detected by the web crawler. The calculation will also consider the weight and normalisation of each parameter.