INFORMATION QUALITY CHALLENGES IN SOCIAL MEDIA

Nitin Agarwal
nxagarwal@ualr.edu

Yusuf Yiliyasi
yxyiliyasi@ualr.edu

Department of Information Science
The University of Arkansas at Little Rock

Abstract

Social media has become an integral part of people's lives. People share their daily activities, experiences, interests, and opinions on social networking websites, opening the floodgates of information that can be analyzed by marketers as well as consumers. However, low barriers to publication and easy-to-use interactive interfaces have contributed to various information quality (IQ) problems in the social media that has made obtaining timely, accurate and relevant information a challenge. Approaches such as data mining and machine learning have only begun to address these challenges. Social media has its own distinct characteristics that warrant specialized approaches. In this paper, we study the unique characteristics of social media and address how existing methods fall short in mitigating the IQ issues it faces. Despite being extensively studied, IQ theories have yet to be embraced in tackling IQ challenges in social media. We redefine social media challenges as IQ challenges. We propose an IQ and Total Data Quality Management (TDQM) approach to the Social media challenges. We map the IQ dimensions, social media categories, social media challenges, and IQ tools in order to bridge the gap between the IQ framework and its application in addressing IQ challenges in social media.

Keywords: Information Quality, Social Media, Blogs, Twitter, Web 2.0, Information Quality Dimensions, Information Overload, Freshness, Relevance, Spam, Splogs, DataFlux, RLab, MatLab, Weka.

INTRODUCTION

Social media, or commonly known as the Social Web, consists of a myriad of services including blogs, social networking websites, wikis, social bookmarking or folksonomies, online media sharing, etc. Through reactive interfaces, low barrier to publication, and zero operational costs, which are all made possible by the new paradigm of Web 2.0, social media has observed a phenomenal growth in user participation leading to a participatory web or citizen journalism. Blogosphere, for instance, has been growing at a phenomenal rate of 100% every 5 months[1]. Technorati has tracked 133 million blogs till December 2008[2]. Other social media sites like Facebook have more than 500 million active users recorded as of September 2010[3]; Twitter amassed nearly

This work is supported in part by Grants from the Office of Naval Research.

[1] http://technorati.com/blogging/state-of-the-blogosphere/

[2] http://thefuturebuzz.com/2008/09/22/technoratis-2008-state-of-the-blogosphere-released/

[3] http://www.facebook.com/press/info.php?statistics

75 million users in January 2010[4]. Other social media sites like Digg, Del.icio.us, Stumbleupon, Flickr, YouTube, etc. are also growing at terrific pace. This clearly shows the popularity of social media among the individuals.

Different social media sites could be alike or different in terms of functionality. Table 1 presents a categorization of various social media sites in terms of functionality. Besides their specific functionalities, they commonly allow individuals to create social connections. Blogs or web logs, is a collection of articles or *blog posts* written by *bloggers* arranged in reverse chronological order. The collection of all the blogs is referred to as Blogosphere. Blogs allow people to share their views, express their opinions, interact and discuss with each other through linking to other blogs or posting comments. Media Sharing sites allow people to annotate, upload, and share their multimedia content on the web, including, images, videos, audio, etc. with other people. Micro Blogging sites are similar to blogs except the constraint on the article length. Twitter[5] allows 140 characters for the posts or *tweets*. These sites are typically used to share what you are doing. Social Bookmarking sites allow people to tag and share their favorite webpages or websites generating community-collaborated metadata for the online media. Social Friendship Networks allow people to stay in touch with their friends and also create new friends. Individuals create their profile on these sites specifying interests, location, education, work, etc. Usually the ties are non-directional, which means that there is a need to reciprocate the friendship relation between two nodes. Social News sites allow people to share, comment, and tag on news with others and let others vote on these stories. Wikis are publicly edited encyclopedias. However, most of the wikis are moderated to protect them from vandalism.

Category	Social Media Sites
Blogs	Wordpress, Blogger, Blogcatalog, MyBlogLog
Media Sharing	Flickr, Photobucket, YouTube, Multiply, Justin.tv, Ustream
Micro Blogging	Twitter, SixApart
Social Bookmarking	Del.icio.us, StumbleUpon
Social Friendship Network	MySpace, Facebook, Friendfeed, Bebo, Orkut, LinkedIn, PatientsLikeMe, DailyStrength
Social News	Digg, Reddit
Wikis	Wikipedia, Wikiversity, Scholarpedia, Ganfyd, AskDrWiki

Table 1: Social Media Sites grouped under Categories based on their Functionalities

Next we will look at the characteristics of social media that makes it an extremely fledgling domain with people not only generating content but also enriching it by providing metadata like tags, labels, categories, etc. Social media sites also provide the capability to be contextually mashed with other sites, generating user-developed widgets. This collaborative environment gives rise to a phenomenon referred to as the *participatory web* or *citizen journalism*.

- **Accessibility** – Social media sites are publicly available for almost free or at no cost. Industrial media is usually privately owned and is not freely available to people.
- **Permanence** – Social media sites can be altered anytime. Individuals can edit their blogs, profile, preferences, etc. anytime. Industrial media cannot be altered once created, e.g., a magazine article that is published cannot be altered instantaneously.
- **Reach** – Like industrial media, social media sites also has a global audience.

[4] http://www.computerworld.com/s/article/9148878/Twitter_now_has_75M_users_most_asleep_at_the_mouse
[5] http://www.twitter.com

- **Recency** – The time lag between communications produced by social media sites can be almost zero. The communication on social media sites can be instantaneous. Whereas, the communications on industrial media can take days, weeks or even months.
- **Usability** – Most social media sites do not require any special skills to create content. Social media sites offer technologies with almost zero operational cost. Whereas, industrial media requires specialized skills and training.

The above characteristics of social media sites lead to various challenges, discussed next.

- **Spam** – Open standards and low barriers to publication have made social media vulnerable to the attacks of spammers. Spammers post non-sensical or gibberish text to social media websites that not only degrades the quality of search results but also consumes valuable network resources. Spam makes it difficult for users to find accurate and relevant information on various social media, discouraging novice users from embracing these technologies and using it as a source of credible information. Bloggers post roughly 80,000 blogs everyday on the internet. Researchers have been debating how many of these blogs are worth reading for useful information, and it is believed that as many as 10,000 of these blogs are in fact marketing hype (Bernstein). Spammers post irrelevant comments on videos at YouTube and on images at Flickr. Due to spammers, user-edited encyclopedia such as Wikipedia has been subjected to vandalism.

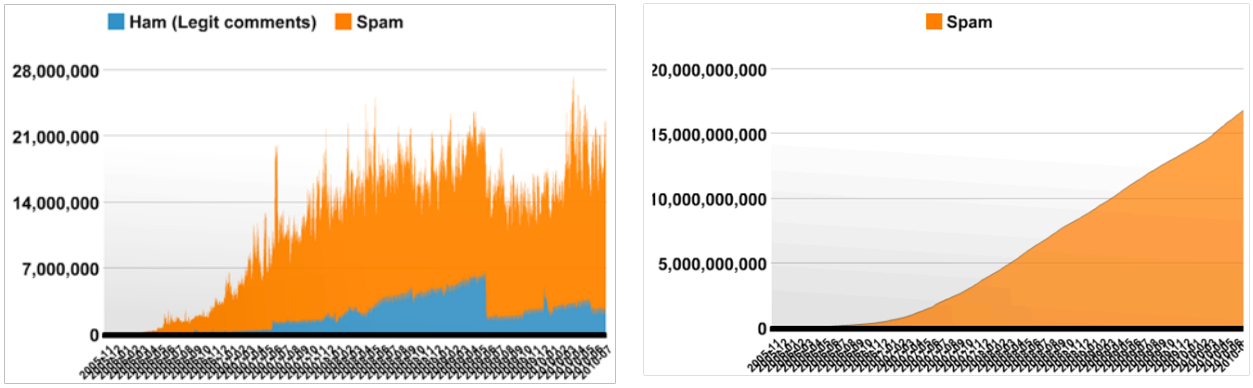

Figure 1: Number of spam comments tracked by Akismet as of July 2010. (a) daily count, and (b) total or cumulative. (Source: http://akismet.com/stats/)

A report published by BBC News in August 2009[6] estimated an astounding 40% spam content on Twitter. One main objective behind this type of spam content is to host content-based advertisements, which would generate revenue if visitors accidentally click on the advertisements. Spammers also create fake blogs or submit spam comments/messages to host link farms, with the purpose of boosting the rank of the participating sites in the link farm (Oard, 2006). Instead of setting up a complex set of webpages that foster a link farm, spammers write a simple agent that visits random blogs, wikis, and media sharing sites and leave comments containing links to the spam content. Often blogs are setup to relay a short message or ping to the servers whenever they are updated with new content. Spinn3r[7] is one such search engine server that receives these pings from blogs to update its search index. In April 2009, Spinn3r reported that they receive nearly 100,000 pings per second from spammers, which forms 93% of the total pings received (http://spinn3r.com/spam-prevention). This indicates the phenomenal growth of spam blogs or *splogs*. In a similar report by Akismet[8] the blogosphere was reported to have 10,085,056,032 total comments till March 2009 out of which only 2,005,536,845 were legitimate

[6] http://news.bbc.co.uk/2/hi/technology/8204842.stm
[7] http://www.spinn3r.com
[8] http://akismet.com/stats/

comments, i.e., over 80% of all comments are spam. The phenomenal growth of spam comments is presented in the Figure 1(a) and (b). Figure 1(a) depicts the daily count of spam and legitimate comments as recorded by Akismet's servers. Figure 1(b) depicts the cumulative or the total number of spam comments received till March 2009, showing an exponential growth in the number of spam comments in the blogosphere.

- **Contextual Relevance** – Information in social media is highly contextual. Information that is relevant to someone may be irrelevant to other(s). For instance, some users on Twitter would like to follow others to be informed of their daily activities. While on the same hand other user(s) may find that annoying.
- **Colloquial Usage and Intentional Misspelling** – Due to the casual nature of the social media, individuals use colloquial forms of language. Such a casual environment nurtures sentiments, expressions, and emotions through writing; it is much more prevalent to observe intentionally modified spellings such as "this is so cooooool…". These instances demonstrate examples of intonation (Prevost, 1996) in written texts. These examples through misspellings clearly emphasize stress on the emotions and convey more information than the regular text. It would be undesirable to disregard them as sheer misspellings. Apparently features that seem noisy might be extremely informative. Services like UrbanDictionary[9] can be used to unravel the informative content in the slangs, abbreviations, and/or colloquial forms of language used by the individuals. Besides colloquial usage, there is a lot of off-topic chatter or noise that could distort the analysis.
- **Information Overload** – With such a rapid pace of content generation as mentioned above, it gets really difficult to follow what is currently happening in social media. The information quickly overwhelms the individuals. Search engines are often faced with the dilemma of choosing freshness of results over accuracy. To address this, one can identify influential individuals (Agarwal, Liu, Tang, & Yu, 2008)and follow them to glean insights of the current affairs.
- **Freshness of Information** – Social media sites are highly dynamic encouraging instantaneous response with almost zero delay in communication. It can be observed from the blogosphere that people have varied interests and their interest in a topic is short-lived (Hayes & Avesani, 2007), causing a drift not only in people's interests but also within the social media.

As the social media continues to grow, the utility of search engines becomes critical. Search engines need to understand the structure and dynamics of the social media to identify the relevant content, and more importantly, to tackle the above-mentioned challenging issues. Researchers have addressed some of these issues that are typical to conventional media such as webpages, emails, etc. from a machine learning perspective (reviewed in next section). We propose to study these challenges under the framework of information quality, i.e., delivering relevant information to appropriate individuals at right time. In this paper, we propose the applications of IQ theory and frameworks by mapping IQ dimensions, social media challenges, social media categories and IQ tools.

In Table 2, we map the social media categories and the social media challenges. For example, some of the major IQ issues in Blogosphere are spam, contextual relevance and freshness of information. Spammers often post spam blogs or irrelevant blogs to distract readers from genuine high quality blogs, or to redirect users to other websites.

[9] http://www.urbandictionary.com

	Spam	Contextual Relevance	Colloquial Usage and Intentional Misspelling	Information Overload	Freshness of Information
Blogs	X	X			X
Media Sharing	X	X			
Micro Blogging	X		X		X
Social Bookmarking	X	X			X
Social Friendship Network			X		
Social News				X	
Wikis				X	

Table 2: Mapping Social Media Categories and Information Quality Challenges

MACHINE LEARNING APPROACHES TO SOCIAL MEDIA CHALLENGES

To address these challenges, supervised machine learning approaches have been studied albeit for rather traditional or conventional media such as webpages and emails. However, the significant differences between the traditional or industrial media and social media (delineated in previous section) warrant a special treatment for information quality aspects in social media. Next we first review existing machine learning techniques using network or link and/or content information followed by the discussion of specialized information quality techniques with their potential to improve and/or complement the existing approaches to handle the information quality issues in social media.

Graph Centric Approaches: Social media can be represented as a network or graph, where the user generated content such as blogs, images, videos, comments, etc., could be treated as nodes and hyperlinks (that cite other such user generated content) as edges. Given such a network of user generated content on social media, we can estimate various statistics like degree distribution, clustering coefficient, and many others that has been shown to differ considerably for spam and non-spam content (Zhu, Sun, & Choi, 2008). These differences could be leveraged to differentiate spam from ham or relevant information from irrelevant. A more sophisticated measure to specially identify splogs leverages the correlation of increment in indegree of blogs and the blog's popularity on a search engine. The assumption here is if a blog is returned among the top search results by a search engine and if it is a legitimate blog, then it will probably attract more inlinks. If the blog was a spam and it somehow managed to get into the top results returned by the search engine through link farming or other tactics, there would be very little or no increase in the inlinks of the splog. To identify spam comments (Kamaliha, Riahi, Qazvinian, & Adibi, 2008) studied network motifs. They identified very peculiar patterns in the commenting behavior of spammers which can be used to further filter out spam comments. Similarly, influential individuals (Agarwal, Liu, Tang, & Yu, 2008) can be identified to address the information overload issues and contextual relevance. However, more work is needed to address various challenges considering the information quality aspects of the social media due to the casual environment.

Content Centric Approaches: Various articles appearing on blog sites can be clustered (Agarwal & Liu, 2009) to provide a summarized view of information hence addressing the information overload. A rather sophisticated approach breaks down the content into different syntactic blocks such as title, tags, comments, content, and links. A separate similarity matrix could be constructed based on segregated information blocks. A combined similarity could be computed using a weighted linear combination of individual similarity matrices (Ntoulas, Najork, Manasse, & Fette, 2006). The problem of identifying spam content can be treated as a binary classification task. User-generated content is assigned one of the two labels: spam or ham. Based on an annotated dataset, a classifier is learned using a portion of the dataset as the training dataset. Remaining portion of the dataset is used as the test dataset to evaluate the efficiency and accuracy of the spam filtering algorithm (Kolari, Java, Finin, & Oates, 2006) (Lin, Sundaram, Chi, Tatemura, & Tseng) (Mishne, Carmel, & Lempel, 2005). Due to the differences between the traditional and the social media mentioned in previous section, existing spam filtering techniques do not perform well. Language models have been studied for traditional media to identify mixture of topics. Topic drift can be studied using longitudinal models comparing different topic distributions identified on a webpage. KL-divergence gives the difference between the two distributions, where Θ_1 and Θ_1 are the two topic distributions. However, often content from social media sites like Twitter could be extremely small (140 characters). This results in a very sparse language model with few words. So in

$$KL(\Theta_1||\Theta_2) = \sum_w p(w|\Theta_1) \log \frac{p(w|\Theta_1)}{p(w|\Theta_2)}$$

order to enrich the model, to achieve more accurate estimation of the language model for both the social media content as well as the comments, links in the content and the comments could be followed and the content found on these links could be added to the existing language models. These links could be followed to a certain depth which would add more content and eventually enrich the language models. Nevertheless, this also leads to the issue of topic drift and hence the language model drifts. Machine learning approaches also suffer from the implicit bias of the training data. These challenges could be handled using established information quality techniques that intend to deliver relevant information to an appropriate audience. Next we investigate some existing information quality approaches and metrics that can be translated as various characteristics of social media. We further analyze the potential of information quality approaches to tackle the challenges faced by social media.

AN IQ APPROACH TO SOCIAL MEDIA CHALLENGES

The Total Quality Management (TQM) methodologies have long been adopted in the traditional manufacturing industries. IQ professionals define information as product of information manufacturing system that process raw data to produce information product (IP) that add value for information consumers (Wang, 1998). IQ professionals have developed an IQ theory and extended the adoption of the TQM methodologies to the world of information products.

Total Data Quality Management

A century of management developments has resulted in the idea of Total Quality Management (Fisher, Lauria, Smith, & Wang, 2008). TQM describes methodologies, concepts, and tools in an effort to promote quality in all operations of an organization (Fisher, Lauria, Smith, & Wang, 2008). Since information can be regarded as product of an information manufacturing system, the TQM approach should therefore also apply to data and

information products. The term Total Data Quality Management (TDQM) is extended from TQM (Fisher, Lauria, Smith, & Wang, 2008). TDQM is an iterative and continuous data quality monitoring and improvement process. We propose the adoption of information quality approach and TDQM methodology to tackle the challenges in social media.

Definitions of Information Quality

In the world of information quality, the terms of information quality and data quality (DQ) are often used interchangeably. Tayi and Ballou define data quality as fitness for use (Tayi & Ballou, 1998). W. E. Deming, who is one of the best known pioneers in the field of quality, stated that "Quality can only be defined in terms of the agent" (Fisher, Lauria, Smith, & Wang, 2008). Wang and Strong adopted TQM approach to data quality and applied advanced statistics to describe correlations among data quality dimensions (Wang, 1998). Four categories of data quality dimensions are proposed as following (Wang, 1998) (Fisher, Lauria, Smith, & Wang, 2008):

Intrinsic IQ reflects the intrinsic nature of data, which means that the quality of the data is knowable only from its use (Fisher, Lauria, Smith, & Wang, 2008).

Contextual IQ means that the quality of data is best determined in the context where it is to be used (Fisher, Lauria, Smith, & Wang, 2008).

Representational IQ describes the presentation and usability of data (Fisher, Lauria, Smith, & Wang, 2008).

Accessibility IQ includes the IQ dimensions of access and security. Access and security reflect the availability of the data as well as its level of protection from unauthorized access (Fisher, Lauria, Smith, & Wang, 2008).

Table 3 summarizes the IQ categories and IQ dimensions.

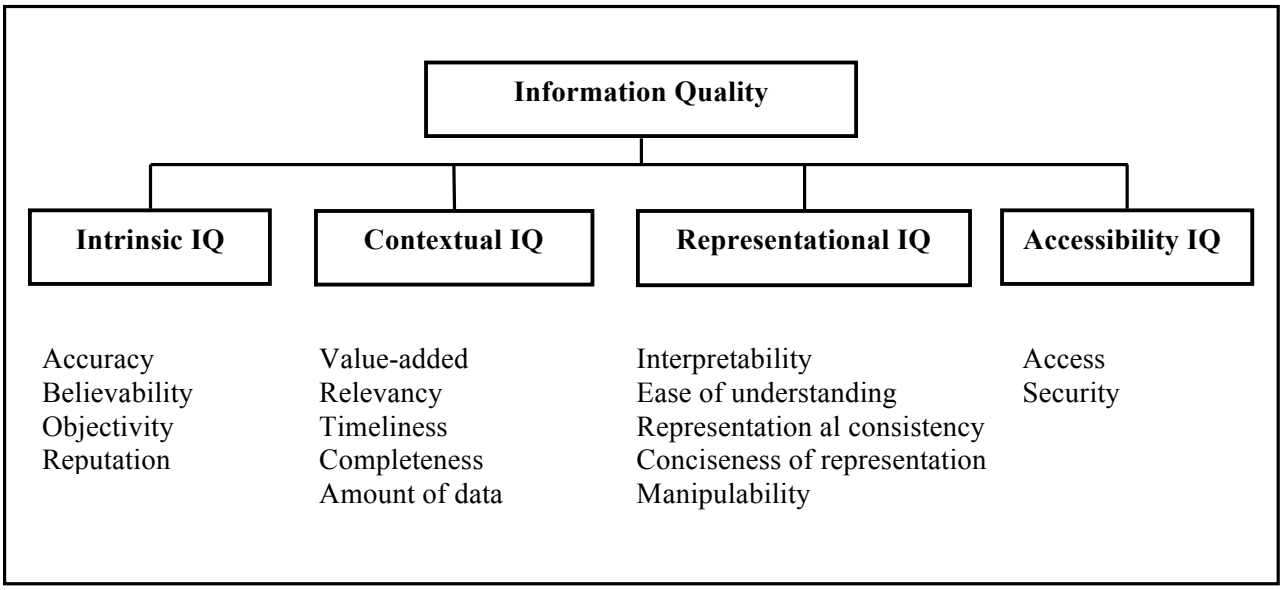

Table 3: Information Quality Categories and Dimensions (Fisher, Lauria, Smith, & Wang, 2008)

In the following, each information quality dimensions are listed and defined in detail.

The Information Quality Dimensions

1. *Accuracy* describes the degree to which data represent the real world. Accuracy is considered to be the most important dimension of data quality (Fisher, Lauria, Smith, & Wang, 2008).
2. *Believability* is the degree to which data is credible. Accurate but unbelievable data is useless for data consumer (Fisher, Lauria, Smith, & Wang, 2008).
3. *Objectivity* describes the degree to which data is impartial (Fisher, Lauria, Smith, & Wang, 2008).
4. *Reputation* of data usually reflects an overall subjective quality assessment of data based on a period of use by its consumers; if data is accurate but lacks good reputation, data consumers might fail to make sound decision using the data (Fisher, Lauria, Smith, & Wang, 2008).
5. *Value-added* reflects the degree to which the use of data delivers benefit to the data consumer (Fisher, Lauria, Smith, & Wang, 2008).
6. *Relevancy* refers the degree to which data is useful for a particular task (Fisher, Lauria, Smith, & Wang, 2008).
7. *Timeliness* of data measures the degree to which the data is current for specific task (Fisher, Lauria, Smith, & Wang, 2008).
8. *Completeness* measures the degree to which data records are available in the information product (Fisher, Lauria, Smith, & Wang, 2008).
9. *Amount of data* refers to the quantity of data. For example, too much information can cause information overload (Fisher, Lauria, Smith, & Wang, 2008).
10. *Interpretability* of data describes the degree to which the data can be presented in a comprehensible manner. Use of unambiguous definitions and vocabularies instead of jargons and acronyms etc. are critical to data consumers' decision making process (Fisher, Lauria, Smith, & Wang, 2008).
11. *Ease of understanding* describes the degree to which data is clear and comprehensible (Fisher, Lauria, Smith, & Wang, 2008).
12. *Consistency* refers to the degree to which the definition, format, and value of data is consistent across systems and applications (Fisher, Lauria, Smith, & Wang, 2008).
13. *Manipulability* describes the degree to which data can be modified, updated, transferred, aggregated, reproduced, integrated and customized in order for it to be utilized in different contexts and for different purposes (Fisher, Lauria, Smith, & Wang, 2008).
14. *Conciseness* describes the degree to which data is brief and to the point. E.g. Long expressions can be difficult to comprehend; on the other hand, short expressions might fail to provide accurate and complete information to its user (Fisher, Lauria, Smith, & Wang, 2008).
15. *Accessibility* describes the degree to which data is available or retrievable. Data accessibility could be inversely related to data security, i.e. high accessibility could compromise data security. On the other hand, excessive data security could decrease data accessibility (Fisher, Lauria, Smith, & Wang, 2008).
16. *Security* is the degree to which data is secure from unauthorized access. As mentioned above, data security could be inversely related to data accessibility (Fisher, Lauria, Smith, & Wang, 2008).

MAPPING IQ DIMENSIONS AND THE CHALLENGES OF SOCIAL MEDIA

In Table 4, the challenges faced by social media are mapped to the IQ dimensions of Wang & Strong framework. These challenges of social media are Spam, Contextual Relevance, Intentional Misspelling, Information Overload, and Freshness of Information. For example, the problems of Spam can be mapped to the IQ dimension of *Reputation* because Spam has low reputation among the users of social media. The Contextual Relevance of social media can be mapped to the IQ dimension of *Relevancy*. Intentional misspelling has low *Accuracy* in terms of spelling even though it might help express the emotions of the user in cases such as "this is coooooooool!". Social media also suffers from the problems of Information Overload which makes decision making more difficult for users. Information Overload can be mapped to the IQ dimension of *Amount of Data*. Last but not least, the Freshness of Information on social media is important to the users as well, for instance, outdated posts can become irrelevant for user needs. Therefore, Freshness of Information can be mapped to the IQ dimension of *Timeliness*. Table 4 summarizes the relationships between social media challenges and IQ dimensions that could be either directly or inversely proportional.

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Completeness	Amount of Data	Interpretability	Ease of Understanding	Consistency	Manipulability	Conciseness	Accessibility	Security
Spam	X	X		X	X	X										
Contextual Relevance						X										
Colloquial Usage and Intentional Misspelling	X				X											
Information Overload									X		X		X	X		
Freshness of Information	X	X		X			X									

Table 4: Mapping Information Quality Challenges and Social Media

MAPPING IQ DIMENSIONS AND THE CATEGORIES OF SOCIAL MEDIA

The IQ dimensions discussed above can be used to assess IQ of social media. In the following, we map the IQ dimensions to the different social media categories as listed in Table 2. Table 5 below shows the significance of individual dimension for different types of social media on a scale of high, medium, and low denoted by H, M,

and L respectively. Not applicable (NA) is used in the case that the IQ dimension does not apply to a specific social media context.

For example, Social News is considered to add little value to the needs of its consumers, and not everyone think that its content is accurate; however, Social News is fast in reporting and can reach its consumers quickly. Therefore, for the scales of significance of Social News, we assign Low, Medium and High respectively to the IQ dimensions of *Value-Added*, *Accuracy* and *Timeliness*.

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Completeness	Amount of Data	Interpretability	Ease of Understanding	Consistency	Manipulability	Conciseness	Accessibility	Security
Blogs	L	M	L	L	L	M	M	L	M	M	H	L	M	L	L	M
Media Sharing	NA	M	L	L	L	M	M	L	M	L	M	L	L	L	M	M
Micro Blogging	L	L	L	L	L	L	H	L	L	L	M	L	L	H	L	L
Social Bookmarking	M	M	L	L	L	M	M	L	L	L	L	L	L	M	L	M
Social Friendship Network	NA	M	L	M	M	L	M	L	L	L	M	M	L	H	H	L
Social News	M	M	M	M	L	M	H	M	L	L	M	M	L	L	H	H
Wikis	M	M	M	M	L	M	L	M	L	H	H	H	H	M	H	M

Table 5: Mapping Information Quality Dimensions and Social Media

Our extensive literature reviews reveal that a wide range of approaches are proposed or adopted by practitioners, academics and IT professionals to address the various information quality challenges in social media, yet, very few have utilized the most current research in IQ theories and best practices. IQ theories have yet to be fully explored and embraced in dealing with the challenges of social media. This research proposes an IQ approach to assess, audit, rank and continuously improve the information quality of social media. State of the art IQ approach can make contributions for this purpose. For example, spams have poor quality in accuracy, relevancy, believability, consistency and reputation etc. Social media with information overload or lack of information have poor quality in the IQ dimension of Amount of Data. A blog post that is outdated has poor quality in the IQ dimension of timeliness. The IQ dimension of consistency is another important criterion for assessing quality of social media. A social media that fail to protect user privacy have poor quality in the IQ dimension of Security. Similarly, a social media that have unreliable accessibility has low quality in the IQ dimension of Accessibility. In the context of any specific social media, the critical IQ dimensions pertaining to its user needs can be identified, ranked and measured for quality. An overall IQ ranking for a specific social media could be obtained by summarizing and weighing the results of each quality dimension measured.

MAPPING INFORMATION QUALITY DIMENSIONS AND INFORMATION QUALITY TOOLS

In Table 6, we map information quality dimensions to the software tools that can help address the information quality issues in corresponding dimensions. For example, DataFlux dfPower Studio is ideal for addressing information quality issues in the quality dimensions of accuracy, completeness, interpretability, consistency, conciseness, accessibility and security.

Information Quality Tools

In the following, we introduce four software tools that can help address information quality issues in social media.

*SAS DataFlux dfPower Studio*is a software tool that allow users to integrate, augment, monitor, and improve data quality throughout an enterprise. The tool allows frontline data clerk to discover and address data problems, merge databases, verify and complete information, transform and standardize data, and perform any other data management tasks necessary.[10]

"*RLab* is a MatLab-like" interactive programming environment, it is a high level programming language intended to provide fast development, as well as data visualization and processing increasing the comprehensibility of data. RLab is a good experimental laboratory ideal for matrix development.[11]

Weka is an open source software tool, it is a collection of machine learning algorithms for data mining tasks. The algorithms can be applied directly to datasets or called from Java code. Weka allows data pre-processing, classification, regression, clustering, association rules analysis, and data visualization. Weka also allows development of new machine learning algorithms.[12]

MatLab, which stands for "Matrix Laboratory", is a computing environment based on fourth-generation programming language. MatLab is ideal for:

- Technical Computing: Mathematical computation, analysis, visualization, and algorithm development
- Embedded Systems: Model, simulate, implement, and verify embedded software and hardware
- Control Systems: Design, test, and implement control systems
- Digital Signal Processing: Analyze signals, develop algorithms, and design DSP systems
- Communications Systems: Design and simulate complex communications systems
- Image and Video Processing: Acquire, process, and analyze images and video for algorithm development and system design
- Test and Measurement: Acquire, analyze, and explore data and automate tests[13]

The results in Table 6 are summarized from our experiments with each of these software tools and the information available on the vendors' websites. In our experiments, we focused on data standardization of addresses and other information using DataFlux. More details are mentioned in Section: Mapping Social Media Categories and Information Quality Tools. Additionally, we also looked at the TREC ClueWeb09[14] dataset with 500 million webpages to extract entities using data mining and social network analysis techniques.

---

[10] http://www.dataflux.com/home.aspx?lang=en-us
[11] http://rlab.sourceforge.net/
[12] http://www.cs.waikato.ac.nz/ml/weka/
[13] http://www.mathworks.com/products/
[14] http://ilps.science.uva.nl/trec-entity/

	Accuracy	Believability	Objectivity	Reputation	Value-added	Relevancy	Timeliness	Completeness	Amount of Data	Interpretability	Ease of Understanding	Consistency	Manipulability	Conciseness	Accessibility	Security
DataFlux dfPower Studio	X							X		X		X		X	X	X
RLab			X								X					
Weka		X			X	X				X	X					
MatLab	X		X		X				X	X	X					

Table 6: Mapping Information Quality Dimensions and Information Quality Tools

MAPPING SOCIAL MEDIA CHALLENGES AND INFORMATION QUALITY TOOLS

In table 7, we map the IQ challenges in social media to the applicable software tools that can help address these challenges. These results can also be obtained directly through Table 4 and Table 6. We have obtained and confirmed these mappings through our experiments with above IQ tools. For example, DataFlux dfPower Studio can help mitigate the information quality issues of colloquial usage and intentional misspelling such as "This is cooooool!". DataFlux dfPower Studio allows data standardization and data cleansing. Any occurrence of the misspelling such as "cooooool" can be detected and replaced with correct spelling. Weka can be used to mine for pattern, correlation, and outlier in social media data; For example, Weka can help detect and monitor IQ issues of spam, contextual relevance and information overload in social media such as blogosphere. Through our analysis, it is understood that various IQ tools could be used in combinations to corroborate findings and to assist decision-making process.

	DataFlux dfPower Studio	RLab	MatLab	Weka
Spam				X
Contextual Relevance	X	X	X	X
Colloquial Usage and Intentional Misspelling	X			X
Information Overload	X			X
Freshness of Information	X			

Table 7: Mapping Social Media Challenges and Information Quality Tools

MAPPING SOCIAL MEDIA CATEGORIES AND INFORMATION QUALITY TOOLS

Table 8 summarizes the mapping of the categories of social media to the applicable software tools. Table 8 can also be derived from Table 5 and Table 6. For example, DataFlux dfPower Studio can be used for the task of standardizing user address data, specified in the profile information at a social friendship networking website (e.g., Myspace, Facebook, etc.), from an optional address format to a preferred address format as shown in the example below. The example can also be extended to include their phone numbers, email address, webpage address, if the physical address is not available.

Optional	*Preferred*
Mr. John Doe CEO Big Business Incorporated Ste 123 987 W Business Ln Kryton Tn 38188	Mr. John Doe CEO Big Business Incorporated 987 W Business Ln Ste 123 Kryton TN 38188-0002

	DataFlux dfPower Studio	RLab	MatLab	Weka
Blogs	X			X
Media Sharing		X	X	X
Micro Blogging	X			X
Social Bookmarking	X			X
Social Friendship Network	X	X	X	X
Social News	X			X
Wikis	X			X

Table 8: Mapping Social Media Categories and Information Quality Tools

CONCLUSIONS AND FUTURE DIRECTIONS

The exponential growth of social media has several far-reaching implications on one hand; yet on the other hand, it suffers from adverse IQ issues. In this paper, we explore several distinct characteristics of social media that makes the IQ challenges it faces very unique. We discuss several IQ challenges in social media. Yet, newer IQ challenges continue to appear every day. For example, intentional misspellings like "this is coooool!" are widely used on social media to embed sentiments with content. State-of-the art approaches are not designed to handle these issues. Methods such as data mining, machine learning, similarity analysis, etc. have only begun to be utilized to address some of these challenges. However, they fall short in taking into account the unique characteristics of social media that are typical of other domains. Despite the advancement and progress in the IQ theories and methodologies, very few researchers and social media operators have focused on or utilized IQ frameworks to address the challenges in the social media. Social media content can be subjectively and objectively measured using IQ metrics and tools. We introduce IQ theories, study multiple IQ dimensions, and introduce several IQ tools. We then map IQ dimensions, social media challenges, social media categories and existing IQ tools. We propose the application of IQ frameworks, TDQM methodologies and IQ tools in tackling the challenges in social media. An IQ and TDQM methodology can help to continuously assess, improve, monitor and audit IQ in social media. When combined with existing industry best practices, our approach has the potential of playing significant role in creating and maintaining high quality social media that deliver high quality information. For future work, we plan to further experiment our approach and develop new IQ tools tailored more specifically for mitigating the IQ challenges in social media.

REFERENCES

[1] Agarwal, N., & Liu, H. (2009). *Modeling and Data Mining in Blogosphere.* (Vol. 1). (R. Grossman, Ed.) Morgan & Claypool Publishers, Synthesis Lectures on Data Mining and Knowledge Discovery.

[2] Agarwal, N., Liu, H., Tang, H., & Yu, P. (2008). Identifying the influential bloggers in a community. *Proceedings of the international Conference on Web Search and Web Data Mining WSDM '08.* Palo Alto, California, USA.

[3] Bernstein, J. (n.d.). *Splogs Serve up Spam to the Blogosphere*. Retrieved November 18, 2009, from www.econtentmag.com

[4] Fisher, C., Lauria, E., Smith, S. C., & Wang, R. (2008). *Introduction to Information Quaity.* MIT Information Quality Program.

[5] Hayes, C., & Avesani, P. (2007). Using tags and clustering to identify topic-relevant blogs. *International Conference on Weblogs and Social Media (ICWSM)*.

[6] Kamaliha, E., Riahi, F., Qazvinian, V., & Adibi, J. (2008). Characterizing network motifs to iden- tify spam comments. *IEEE International Conference on Data Mining Workshops, 2008. ICDMW '08*, (pp. 919–928).

[7] Kolari, P., Java, A., Finin, T., & Oates, T. (2006). Detecting Spam Blogs: A Machine Learning Approach. American Association for Artificial Intelligence.

[8] Lin, Y., Sundaram, H., Chi, Y., Tatemura, J., & Tseng, B. (n.d.). Detecting splogs via temporal dynamics using self-similarity analysis.

[9] Mishne, G., Carmel, D., & Lempel, R. (2005). Blocking blog spam with language model disagreement. *First International Workshop on Adversarial Information Retrieval on the Web (AIRWeb)*.

[10] Ntoulas, A., Najork, M., Manasse, M., & Fette, D. (2006). Detecting spam web pages through content analysis. *15th international conference on World Wide Web.*

[11] Oard, D. W. (2006). Splog! Or how to stop the rise of a new menace on the internet. *Harvard Journal of Law & Technology , 19*.

[12] Prevost, S. (1996). An information structural approach to spoken language generation. *Proceedings of the 34th Annual Meeting on Association For Computational Linguistics.* Santa Cruz, California.

[13] Tayi, G., & Ballou, D. (1998). Examining Data Quality. *Communications of the ACM , 41* (2), 54-57.

[14] Wang, R. Y. (1998). A Product Pespective on Total Data Quality Managment. *Communications of the ACM , 41* (2).

[15] Zhu, L., Sun, A., & Choi, B. (2008). Online spam-blog detection through blog search. *Proceedings of the Seventeenth ACM International Conference on Information and Knowledge Management (CIKM)*, (pp. 1347–1348).