(304) Guidelines for Accessibility to Microblogging: an Integral Approach

Yod Samuel Martín García Laura Álvarez Alba Juan Carlos Yelmo García
Universidad Politécnica de Madrid – Departamento de Ingeniería de Sistemas Telemáticos
{samuelm, lalvarez, jcyelmo}@dit.upm.es

Abstract

Microblogging is one of the most popular user-generated media (UGC) types; hence its accessibility has a large impact for users. However, the accessibility of this medium is poor actually, due to the combination of bad practices by different agents, ranging from the providers that host microblogging services to the prosumers that post contents to them. Here we present a model introducing the different components that play a part in microblogging services from the perspective of accessibility; then we analyze the impact of each of them and propose some guidelines so that they may meet accessibility requirements. In particular, we base on a study performed on Twitter (one of the most relevant microblogging platforms) to identify good and bad practices regarding accessibility in microblogging content generation.

Resumen

Los 'microblogs' son uno de los tipos más populares de contenido generado por los usuarios (UGC), por lo que su accesibilidad puede tener un gran impacto. Sin embargo, en realidad este medio ofrece una accesibilidad muy pobre por la combinación de malas prácticas, cuyo origen va desde los proveedores que albergan los servicios de microblogging hasta los 'prosumidores' que envían los contenidos. En la presente ponencia, se presenta un modelo de los distintos componentes de los servicios de microblogging desde el punto de vista de la accesibilidad, se analiza el impacto de cada uno de ellos, y se proponen algunas pautas para que cumplan con los requisitos de accesibilidad. En concreto, realizamos un estudio sobre Twitter –una de las plataformas más relevantes de microblogging– para identificar buenas y malas prácticas de accesibilidad en la generación de contenidos de microblogging.

1. Introduction

Web users are not mere passive content consumers any more, but they have also become active contributors in the Web 2.0 sites. This has led to the dual role labeled as "prosumer", which depicts users who both consume contents created by others and produce their own ones, thus engaging into a communal creation process. User-Generated Contents (UGC) –defined as those publicly available contents created as a result of creative effort by non-professional users [1]– are becoming more and more widespread, up to the point that almost 1 out of 3 web sites in the top 1000 (as measured from Alexa [2]) offer UGC as a relevant part of their contents.

However, this large dissemination does not usually go by high quality standards –and that also affects accessibility. UGC have their specific accessibility problems [3]: on one hand, prosumers are neither trained on, nor acquainted with, or aware of accessibility issues; on the other, they do not have to respond to clients, thus lacking any accountability and devoting little resources to improve the quality of those often short-lived contents.

In the rest of the paper we present a study on the accessibility of microblogging –an especially relevant type of UGC–, and outline some possible techniques to improve it. Section 2 introduces the different components present in the usual microblogging scenario regarding their role for accessibility, which we deal with in the next sections. Section 3 defines possible approaches for the platforms and the user-agents to improve accessibility to microblogging contents. Section 4 presents the user practices observed in a massive study performed on Twitter (the most used microblogging service); from which we extract some guidelines for users in section 5. Finally, section 6 concludes the article and presents prospective future research lines.

2. Microblogging services and contents from the perspective of accessibility

2.1. RELEVANCE OF MICROBLOGGING

Microblogging [4][5] is a service that allows users publish on the Internet small elements of content. Same as it happens on fully-featured blogs, microblogging topics range from casual, personal matters, to hobbies or marketing and promotion from brands or firms. Microblogging services usually allow users to subscribe to contents published by others, so that they may check them on real-time from the service user interface.

The most used microblogging service is Twitter, yet it coexists with others such as Tumblr, or even the so called "status update" services by online social networks (such as Facebook Wall, Yahoo Pulse, Google Buzz, etc.) All these services are proprietary, in that they do not allow users from one service to subscribe to feeds hosted by another one. On the other

hand, there are commercial and open-source products, such as OStatus, that allow organizations to set up their own microblogging services and interoperate with one another, either for corporate use or as a service provider for external users (e.g. Identi.ca, Status.net).

Microblogging is reaching a large degree of social influence [6]. Twitter, the most relevant microblogging service, currently hosts more than 50 billion entries, growing exponentially with 1 billion more currently being added each week. It is used by companies as a way to be in touch with their customers and swiftly diffuse their messages and campaigns, taking advantage of the so-called "viral marketing". It has been pivotal for the self-organization in the popular upheavals that have been recently developing in North African and Arab countries. In conclusion, microblogging services are a powerful communication tool with a large social relevance nowadays, and accessibility barriers in those services would preclude many users from a full involvement and participation in the society.

2.2. IMPACT OF THE COMPONENTS OF A MICROBLOGGING SERVICE ON ACCESSIBILITY

In order to understand the good and bad practices regarding accessibility in microblogging services, and the best ways to address them, we introduce a model that shows all the components taking part in the workflow of a typical microblogging scenario, and their relation with accessibility. We have compiled this model based on Twitter, yet it can be easily adapted to any other microblogging service. This model integrates two different viewpoints: the components of web accessibility as defined by the WAI (Web Accessibility Initiative) [7] together with the usual Model-View-Controller [8] and 3-tier client/server architectures [9] typical of web applications. Following we detail the role of each of the agents participating in the workflow, shown in Figure 1 on next page.

2.2.1. Content producers. As above explained, the authors are usually non-professional creators, with the implications that entails for accessibility. Several techniques (documentation, guidance, etc.) may be employed in order to promote the creation of accessible contents among the producers.

One of the most salient features of microblogging users is the communal generation of a consensus for the language, model and processes employed. The tight limits in the brevity of contents has forced users to devise new ways to add deep meanings in just a few characters, and thus has given rise to new syntactic conventions. For instance, a hash sign ("#") is prepended to terms referencing common topics, a caret ("^") to author signatures, a commercial-at sign ("@") to user mentions, etc. This is also shown in the language employed, where colloquial or ad hoc abbreviations are commonly used to condensate many ideas in such a short space.

2.2.2. Content editor. The users create their microblogging posts, called "tweets", using different content editors, which play the role of *authoring tools*. Twitter itself provides its own, plain editor on its web site, but other third

parties also provide web-, mobile- or desktop-based applications to create and post new contents to Twitter. Finally, users also employ the editors to manage the service (for subscriptions, configuration, etc.) Apart from the traditional, user-driven editors, any authorized service may auto-generate and post tweets without any user intervention (e.g., to send alerts triggered by an external event, etc.)

2.2.3. Ancillary storage services. Since Twitter only provides a limited capacity for each tweet, external services have arisen that allow users to create, upload, or link additional contents that will be hosted on external services and linked from the original tweet. For instance, there are external image-hosting services specifically designed to have them linked from Twitter. However, the most paradigmatic example is the rise of URL shorteners: services that just provide a redirection facility from a URL a few characters long, to a destination website elsewhere.

2.2.4. Semantic data model and business logic. Twitter hosts the tweets in a database system where they are stored together with related semantic information. Tweets themselves just consist of 140 characters at most, but Twitter does not store them in a plain format, but decorated with several kinds of semantic information:

On the one hand, we have *extrinsic metadata* pertaining the tweet, such as its author, creation date, original source (in case it was originated by forwarding or replying to a different tweet), or geo-location.

On the other, a tweet can be enriched with annotations that describe some parts of its contents, which Twitter has integrated mimicking those community uses above presented. Thus, if a tweet contains a URL, a mention to a Twitter user, a reference to a common topic (called “hashtag”), or the signature of an individual author contributing in a collectivity (or “cotag”); then an annotation is stored together with the tweet signaling the special semantics of that part of the text.

Even more, Twitter is capable of identifying the usage of some external storage services and taking that into account for the annotations. On top of that, Twitter offers a framework to provide ad hoc annotations of any user-defined types (yet Twitter administrators themselves suggest some possible use schemes). All this information is accessed through a standardized API (Application Programming Interface) [10], where external clients can post or retrieve tweets with all the semantic information needed.

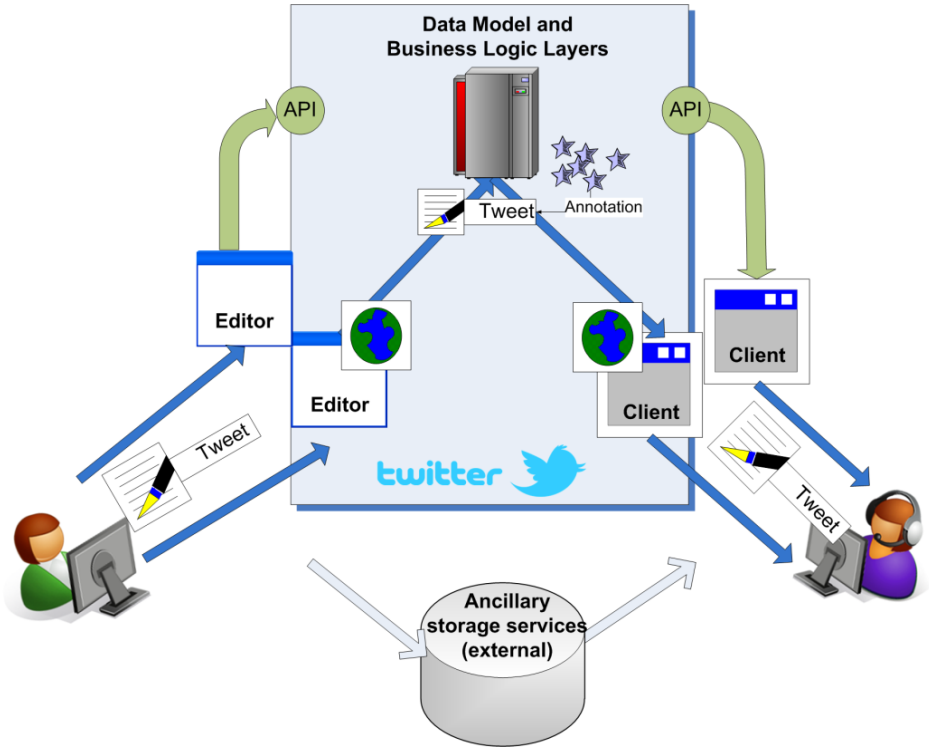

Figure 1. Components of a microblogging service.

2.2.5. Presentation or client layer. The presentation layer translates the data stored by Twitter into a user-interface definition (e.g. a Web page) that presents data in a human-readable fashion; thus, it plays both the role of an *authoring tool* (it is in charge of creating content) and a *user agent* (it provides the user interface).

Twitter itself provides users with a Web-based service to access tweets based on two different services: it hosts a public page for each user, where their tweets are published and may be accessed by anyone; and it provides as well a personalized "timeline" page where users may get the last tweets published by all the users they are subscribed to. These pages have their specific presentation features: users may customize the colors and background image of their public pages; while Twitter defines a fixed template for the page structure, font faces, etc.

However, other tools also exist to read Twitter, based on the public APIs that expose the data, as just presented. These tools may share or not the same presentation features as those provided by the Twitter website, and they may use completely different interface technologies (e.g. a desktop application) or modalities (e.g. speech synthesis to read the contents). Even though tweets are formally nothing more than 140 characters, all these tools may provide more information based on the tweet metadata and annotations. As simple examples, URLs can be marked as links in HTML, or the externally stored images linked from a tweet can be rendered together with the tweet, etc.

3. Transversal solutions for accessible microblogging

As we have explained in the previous section, there are several agents contributing to the contents perceived by the final user –each of them having its own impact on accessibility. The creator is responsible in part of the accessibility of the content he or she creates, but we should not dismiss the role of the other agents: here we explain the effects they may produce on accessibility.

3.1. PLATFORM-BASED SOLUTIONS

The major contributions to accessibility by the platform that stores and processes the contents may come from two approaches: a *richer metadata model*, and *semantic preservation and augmentation*.

Most of the accessibility problems found in microblogging platforms come from the medium constraints: a short string of plain text. However, nothing should preclude users from providing additional hidden data with information for accessibility, without needing to overflow the size of the message eventually rendered to the user. It has been conjectured that these metadata constitute the natural evolution of microblogging [11], and Twitter itself has been adding several metadata items to its data model for tweets. For instance, labeling the natural language of an individual tweet or part of it does not increase its practical length, but it adds a much needed piece of metadata that screen readers may use as a hint for pronunciation. This applies, in general, to any markup that could be added to ease accessibility (titles, acronym expansions, quotations)

The second approach implies an active task by the platform. Aside from letting users add more metadata to content, the platform should always preserve it and even add more on their own. For instance, it could recognize URLs, emoticons, etc., and label them properly. Moreover, this active task can be extended to provide guidance to the creators: e.g. disallowing inaccessible color combinations, using face-recognition software to assess the adequacy of the profile photograph, precluding users from sharing links that do not have any explaining text, etc.

3.2. USER-AGENT-BASED SOLUTIONS

At the other side of the process, we find the different presentation tools. Their main role regarding accessibility is that of providing access to any piece of information available regarding the microblogging post, be it part of its content, its metadata, or data stored by an ancillary service. Thus, a high-quality user-agent, would present:

All the annotated entities with a distinct presentation (e.g. links underlined, quotations rendered between quotes or uttered with a different voice), skipping out unnecessary conventions (e.g. extraneous signs).

All the contents obeying the preferences dictated by the consuming user, in order to avoid any potential problem coming from an incorrect design (contrasts, etc.)

All the metadata available from each microblogging post, wherever it might be stored. This includes, e.g. the author's avatar, the images linked from the post and their alternative text, the title of the destination page of a link (resolving all the redirections if needed), etc.

As user agents such as EasyChirp[6] or Syrinx have proved, the presentation of tweets does not need to be inaccessible –it may rather be as accessible as the developer of the user agent wants.

4. Field study on the accessibility of microblogging contents

In order to determine the impact of accessibility issues of UGC in microblogging, we have developed a field study over a broad set of Twitter contents.

4.1. SCOPE AND TARGET

Using Twitter's API, we mined Twitter to retrieve a broad set of contents that could provide a representative sample. For that, we have followed several, complementary strategies, retrieving:

Random tweets, at a rate of 20 per minute during one week.

Popular tweets, either being relevant on their own as identified by Twitter's API, or pertaining to globally relevant discussion topics (called "trending topics").

Tweets from popular users (usually celebrities, bloggers or corporations), as identified by Twittercounter [12] statistics service. We should note that these types of users generate most of the impact in Twitter [13].

In order to automate the evaluation of the results over such a large sample, we have restrained to the evaluation of a limited subset of accessibility criteria: vocabulary (encompassing language clarity, abbreviations, etc.), link significance, metadata, and design. We refer the readers to accessibility guideline families [14] to check how each of these aspects in particular affects accessibility.

4.2. VOCABULARY

As we have explained, the community of Twitter users has created its own linguistic codes, which sometimes depart from the conventional usage. If users find terms that are not part of their natural language, they may encounter serious accessibility problems:

Screen reading software will not correctly read those non-lexical tokens, or it will generate awkward utterances (or just gibberish).

[6] Formerly AccessibleTwitter

Users with limited reading competences or dyslexia will be confused by the language employed and not be able to understand the contents, etc.

There are several issues that fall under this category:

Usage of specific symbolic characters prepended to, appended to, or enclosing a term to denote a special meaning (hash for topics, caret for signatures, commercial-at for user mentions, etc.)

Groupings of words in a single token without blank spaces, to denote specific entities, in combination with the techniques just mentioned.

Usage of symbolic or non-Latin Unicode characters that exhibit a visual resemblance to their Latin counterparts, in order to create decorative text (e.g. the lowercase Greek letter eta "η" for the Latin "n").

Usage of colloquial abbreviations or ad hoc spellings that reduce the number of characters (e.g. the letter "u" or the number "4" respectively standing for the pronoun "you" or the preposition "for").

Usage of iconic characters (dingbats) to transmit concepts in a condensed way (e.g. a heart character "♥" to mean "love").

Usage of URLs as the text of links, since they do not follow natural language rules; especially when they are pointing to a URL-shortening service, which hides any hint that the original URL could have provided under an obfuscated alphanumeric string.

Usage of natural languages different from that declared for the tweet.

Several of these may appear combined together, e.g., a user may write "#ff @jsmith" to signal "today Friday, I recommend subscribing to the contents of the user John Smith". Even though these problems have different origins and solutions, all show as words that are not recognized as part of the target natural language, which allows us treating all of them together.

In order to analyze the impact of the vocabulary used on the accessibility of the contents, we have followed the following procedure for each tweet analyzed:

Select tweets in English or Spanish (for which we possess morphologic analysis tools).

"*Whiten*" each tweet, removing all the annotated entities, based on the available metadata. These entities are deemed as tokens that never pertain to the vocabulary of the language.

Lemmatize the contents of each tweet, that is, split it into words and reduce each to its base form (without any morphological flexions). We have leveraged on Freeling morphological analysis tool [15] for that process.

Compute the self-information of each word, measured in *bits* as given by the following definition of self-information: $I(w_i) = -\log_2(p_i)$, where p represents the probability of each word to appear in that natural language. The probabilities have been drawn from the frequencies in the corpora developed by University of Leeds [16], and a reasonable value has been estimated following Zipf's law for those terms not appearing in the corpus.

Add this quantity up over all the words contained in a tweet, thus obtaining the self-information of the whole tweet (supposing statistical independence between words), and *compare* that with the average entropy of the respective natural language:

$$\sum_{w_i \in \text{tweet}} -\log_2(p_i) - \sum_{w_i \in \text{Corpus}} -p_i \cdot \log_2(p_i)$$

This entropy analysis above explained, yielded *more than 100 bits of excess information on average of each tweet above the expectation*. In summary, this means large readability problems, due to any of the issues explained at the beginning of this subsection.

4.3. LINK SIGNIFICANCE

It is important that the text of a link clearly identifies its target. For those users sequentially navigating through a list of links, the text should be clear enough that its target can be distinguished in isolation. If not, at least the text surrounding the link in the same paragraph should help identify the target

However, link texts in microblogging services usually consist of the URL itself (which is not relevant enough at all), or even the URL of a redirection service, which even precludes the user from figuring out the destination site (thus being exposed to possible scams, undesired content, etc.)

In addition, links to tweets related to common topics (hashtags) are not used in a consistent way, since they are created by the community, and different users may be using the same hashtag with different meanings of vice versa.

4.4. METADATA AND SEMANTIC ANNOTATIONS

Tweets may include semantic annotations, which may help overcome the limitations imposed by the 140-character limit and include much more useful information to improve the accessibility of the contents. For instance, if part of the content is identified as a URL, a microblogging user-agent could well present the title of the document identified by the URL instead of the sequence of characters that make it.

We have thus evaluated the appearance of several metadata types in tweet structures. Following this analysis, we found a per-tweet average of 0.2 hashtags, 0.25 URLs (only 8% of which provided an expanded URL to display in replacement of a shortened one) and 0.37 user mentions. All of them add to the entropy surplus presented in the previous subsection.

4.5. DESIGN

Even though Twitter establishes the main design of a user's page, there are several elements the user may customize. We may take into account at least three aspects:

User avatar: a small image appears on top of each user's page, as well as together with each tweet by him or her elsewhere included. This image must correctly represent the user: e.g. it must be a photograph of that user's face, with good lighting conditions, contrast, etc. This will be helpful for people with cognitive impairments or low vision, in order to identify the referred user. Alternative text is of course

also relevant, but it cannot be currently defined by the user in Twitter: we advise providing a proper user description in the field devoted for that.

Background image: in order to overcome the design limitations imposed by Twitter, many users include their own texts embedded in the background image of their page. Needless to say these texts will not be accessible for anyone who is not accessing the contents through a graphical user interface.

Color combination: Twitter allows users to customize the foreground and background colors of the different elements of their page. If color, contrast and brightness ratios between elements are not enough, they will pose accessibility problems to people with low vision or color blindness.

Regarding the use of design templates, we found that:

the majority of users employed the default combination provided by Twitter (thus not introducing any additional accessibility problems);

they did not use personal photographs as user avatars (difficulting recognition); and

the presence of semantic annotations was testimonial (a few cases in more than $2 \cdot 10^5$ tweets).

5. User-oriented guidelines for generating accessible microblogging contents

Based on the practices observed in the study described in the previous section, we have compiled a set of guidelines targeting the creators of microblogging contents:

Use a profile picture where the user appears in the foreground, without anybody else, and with sufficient contrast.

Fill in all the metadata fields available when posting some content.

Do not embed texts in the background images of the user page.

Avoid emoticons, “leet-speak”, fancy characters, or any other kind of text whose intended meaning relies on a specific visual presentation.

Use contrasting font and foreground colors, choosing preferably the default combinations.

Use the tools provided by the editor to mark quotations and links.

Avoid colloquial and shorthand abbreviations. Use concise language and less verbose wordings instead, or transmit fewer ideas. Exhaust all the available length of a micropost to avoid unnecessary abbreviations.

Use the clearest possible language.

6. Conclusions and future work

Here we have presented several approaches to improve the accessibility of user-generated contents, specifically addressing microblogging. In any case, they must encompass all the agents involved in the workflow of microblogging production to ensure real accessibility for the end-user.

We aim at continuing our research with deeper mining and analysis of the results collected, dealing with specific accessibility checkpoints. In addition, we plan to expand the research to other microblogging services and alike, for which an open API exists (such as Facebook status service, Google Buzz or OStatus).

7. References

[1] Participative Web and User-Created Content: Web 2.0, Wikis and Social Networking. SourceOCDE Science et technologies de l'information, 2007, 15 (Sep. 2007). OECD Organisation for Economic Co-operation and Development.

[2] Alexa Top 500 Sites [http://www.alexa.com/site/ds/top_sites]

[3] Y.S. Martín García, B. San Miguel González y J.C. Yelmo García. Prosumers and accessibility: how to ensure a productive interaction. In Proceedings of the 2009 International Cross-Disciplinary Conference on Web Accessibililty (W4A) (W4A '09), 2009. ACM, New York, NY, USA, 50-53.

[4] A. M. Kaplan, M. Haenlein, The early bird catches the news: Nine things you should know about micro-blogging, Business Horizons, Volume 54, Issue 2, March-April 2011, Pages 105-113.

[5] P. McFedries,; , "Technically Speaking: All A-Twitter," Spectrum, IEEE , vol.44, no.10, pp.84, Oct. 2007

[6] A. Java, X. Song, T. Finin, and B. Tseng. 2007. Why we twitter: understanding microblogging usage and communities. In Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis (WebKDD/SNA-KDD '07). ACM, New York, NY, USA, 56-65.

[7] Chisholm, W. A. and Henry, S. L. 2005. Interdependent components of web accessibility. In Proceedings of the 2005 international Cross-Disciplinary Workshop on Web Accessibility (W4a) (Chiba, Japan, May 10 - 10, 2005). W4A '05, vol. 88. ACM, New York, NY, 31-37.

[8] Avraham Leff and James T. Rayfield. 2001. Web-Application Development Using the Model/View/Controller Design Pattern. In Proceedings of the 5th IEEE International Conference on Enterprise Distributed Object Computing (EDOC '01). IEEE Computer Society, Washington, DC, USA, 118-.

[9] W. W. Eckerson. Three tier Client/Server architectures: Achieving scalability, performance, and efficiency in Client/Server applications. Open Information Systems, 3(20):46–50, 1995.

[10] Twitter. API Documentation [http://dev.twitter.com/doc]

[11] John Breslin and Stefan Decker. 2007. The Future of Social Networks on the Internet: The Need for Semantics. IEEE Internet Computing 11, 6 (November 2007), 86-90.

[12] Twitter Counter [http://www.twittercounter.com]

[13] Wu, S.; Hofman, J.M.; Mason, W.A.; Watts, D.J. Who Says What to Whom on Twitter. Yahoo Research, 2011. [http://research.yahoo.com/pub/3386]

[14] B. Caldwell, M. Cooper, L. G. Reid, G. Vanderheiden. Web Content Accessibility Guidelines (WCAG) 2.0 W3C Recommendation 11 December 2008.

[15] Lluís Padró and Miquel Collado and Samuel Reese and Marina Lloberes and Irene Castellón. FreeLing 2.1: Five Years of Open-Source Language Processing Tools. Proceedings of 7th Language Resources and Evaluation Conference (LREC 2010), ELRA. La Valletta, Malta. May, 2010.

[16] Large Corpora. Centre for Translation Studies, University of Leeds. [http://corpus.leeds.ac.uk/list.html]