Producing Volunteered Geographic Information from Social Media for LBSN Improvement

Maxwell Guimarães de Oliveira, Cláudio de Souza Baptista, Cláudio E. C. Campelo, José Amilton Moura Acioli Filho, Ana Gabrielle Ramos Falcão

Laboratório de Sistemas de Informação, Universidade Federal de Campina Grande (UFCG) - Brazil
maxwell@ufcg.edu.br, {baptista, campelo}@dsc.ufcg.edu.br,
{joseamilton, anagabrielle}@copin.ufcg.edu.br

Abstract. Volunteered Geographic Information (VGI) emerged from the widespread of devices featuring GPS and Internet connectivity around the world. It has enabled the easier and increased production of spatial data, and a deeper engagement of people with everything involving location. Such scenario has led to the emergence of Location-Based Social Networks (LBSN), which allow users to be assigned to space related content. LBSN environments have proved to be quite useful, however, keeping users willing to contribute (i.e., maintaining such environments in continuous operation) has appeared to be challenging. In addressing this issue, we have considered applying Geographical Information Retrieval (GIR) techniques to produce VGI from social media streams on the Web, aiming to improve LBSN with valuable up-to-date content in an automated way. We rely on GIR techniques such as geoparsing the message bodies instead of considering previously geotagged information since we cannot ensure that an embedded geolocation is the same location that the messages refers to. An artifact for automatically producing VGI based on social media content is described and validated using a real-world case study. We harvested tweets during the FIFA Confederations Cup and tried to produce valuable VGI from the message stream. Our results proved to be promising for leveraging VGI from social media.

Categories and Subject Descriptors: H.2.8 [**Database Applications**]: Spatial databases and GIS; H.3 [**Information Storage and Retrieval**]: Miscellaneous; I.7 [**Document and Text Processing**]: Miscellaneous

Keywords: Geoparsing, GIR, LBSN, Twitter, VGI

1. INTRODUCTION

The easier and increased production of spatial data has enabled a deeper engagement of people with everything involving location. It can be mainly explained by the dissemination of devices featuring GPS and the spread of Internet connectivity around the world. Space related information have been shared and also consumed by thousands of users by means of location-based applications. The OpenStreetMap, for example, has over 2 million users, over 8 billion records and over 10 thousand data updates per week[1]. On the other hand, a large number of users have been increasingly consuming such information by means of location-based applications.

Such phenomenon enabled the emergence of Volunteered Geographic Information (VGI) as an alternative and powerful spatial data source on the Web. VGI consists of geographic data provided voluntarily by individuals who act as sensors [Goodchild 2007]. Since the main feature of the Crowdsourcing phenomenon [Surowiecki 2005] is the dissemination of data produced by people spread around the world, VGI emerges as a specific kind of crowdsourced data. Most of these volunteers are ordinary people interested in sharing their footprints, viewpoints and knowledge about geographic locations.

[1]http://www.openstreetmap.org/stats/data_stats.html (last access in 15/07/2015)

Research supported by the *Conselho Nacional de Desenvolvimento Científico e Tecnológico*, CNPq, Brazil.

This scenario has led to the emergence of applications such as the Location-Based Social Networks (LBSN). LBSN [Vicente et al. 2011] comprises crowdsourcing environments that have enabled users to share valuable spatial, temporal and semantic information in a specific knowledge domain, such as entertainment and public utilities. The Crowd4City, for instance, is a LBSN that can be applied to the domain of smart cities, with support for participatory human sensors [Falcão et al. 2012]. Crowd4City aims at creating an environment for identification and discussion of matters concerning the government of the cities, a common interest of the population.

Despite initiatives like Crowd4City, one of the main challenges in the use of human sensors has been keeping their willingness to cooperate and consequently maintaining the LBSNs continuously. Several factors may affect the users' motivation, such as the learning curve for correct operation of a LBSN and the time spent on such activities. Typically, only a few users are in charge of providing a significant volume of information. This issue is visible in terms of geographic location, where many areas around the world are mapped by only few users [Haklay and Weber 2008]. Therefore, it becomes necessary to find alternative methods of keeping the LBSNs up-to-date even when the volume of contributions from volunteers is below the expected number. In addressing this issue, we have considered applying Geographic Information Retrieval (GIR) techniques to automatically produce VGI for LBSN improvement. We understand that produced spatiotemporal markers conforms to VGI policy as they are produced by people as the same as VGI is produced, and share features of spatial crowdsourced data. Thus, social media users non-intentionally will be in charge of volunteers in the automated production of VGI.

One of the sub-areas of Geographic Information Retrieval (GIR) focus on the development of techniques for identifying geographic locations associated with text documents. Research in this area poses many challenges, including those relating to Natural Language Processing (NLP), handling of uncertainty, disambiguation, and context identification [Bordogna and andGiuseppe Psaila 2012]. Through these GIR techniques, it is possible to process text (from websites, blogs and social networks, for example) and then assign specific geographic locations to it [Purves and Jones 2011]. In this sense, previous work has addressed the assignment of geographic locations to Web documents, including social network messages such as tweets from Twitter[2] [Rupp et al. 2013; Watanabe et al. 2011].

Our work builds on the hypothesis that texts from the Web, such as messages publicly exchanged in social networks, when processed by a mechanism for identifying the geographic locations they are associated with, could automatically turn into useful information to feed location-based applications such as LBSNs. Hence, we consider that geographic information originated from social media streams can also be classified as VGI. Thus, social media producers assume the role of volunteers in the production of VGI, which could automatically be made available to users of LBSN applications.

In order to validate this hypothesis, this article presents an approach to automatically producing VGI from social media streams for enriching LBSNs. This approach is based on the application of GIR techniques to text messages from social media. Although public data originated from social media already contain some geotagged information in the form of metadata (such as the geographic position of the user who posted a message), users can freely disseminate information about the most diversified geographic contexts, which frequently mismatch their geographic position at the moment that the information is shared. Thus, by generating VGI from a text message, we can ensure the identified locations actually relate to the places the message content refers to. In our envisioned scenario, the produced VGI should become available for LBSN users, who will be the main consumers and also validators of that information, being capable of pointing to its inconsistencies as well as stressing its relevance, and consequently enriching the crowdsourcing environment.

The main contributions of this article are: the description of a mechanism developed for automatically producing Volunteered Geographic Information, based on the content of social media texts; and

[2]Twitter: http://www.twitter.com

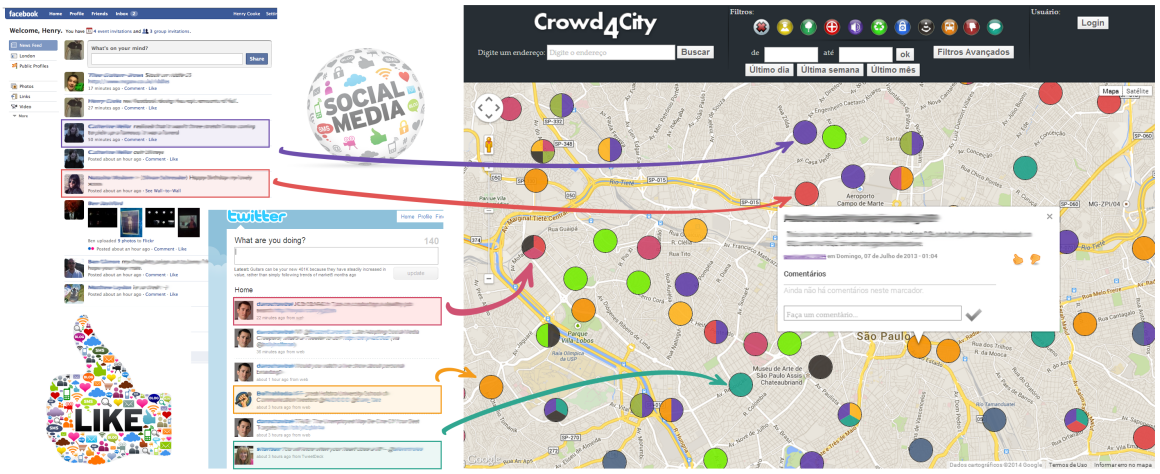

Fig. 1. The main idea of our proposal: turning social media messages into valuable VGI in a LBSN

a discussion on the task of geoparsing informal texts from social media and how valuable the information can become when mentioned locations are explored. The remainder of this article is structured as follows. Section 2 describes our proposal. Section 3 addresses a case study carried out to evaluate the proposed ideas. Section 4 discusses related work. Finally, section 5 summarizes and highlights further work to be undertaken.

2. AN APPROACH FOR AUTOMATED PRODUCTION OF VGI FROM SOCIAL MEDIA

This section presents our approach to automated production of VGI based on crowdsourced data from social media. We propose a systematic approach aiming at the automatic production of VGI based on information published on Twitter microtexts. The expected result is the production of spatiotemporal markers with the content of these microtexts, which can be widely viewed and handled by the users of a LBSN. We have started the designing and implementation of our approach focusing solely on Twitter.

We believe the automatically produced VGI may help LBSN users interested in learning more about specific geographical locations from people who freely share valuable information on social media. An overview of the proposed approach is presented in Figure 1.

Figure 1 (left side) illustrates the social networks, such as Twitter and Facebook[3], as information sources for producing the VGI visualized in a LBSN such as the Crowd4City (right side). In this context, each message posted by the users of these networks can be turned into VGI, which can then be used by LSBN users.

In order to achieve this goal, it is necessary to have a computational process that involves the capture and the treatment of information using GIR techniques, as illustrated in Figure 2. The computational process of this approach involves, basically, four distinct stages: Crawling, Geoparsing, Georeferencing and VGI Production. The initial stage is Crawling, in which the messages posted on the social networks are captured. We developed an algorithm to capture real-time microtexts (tweets) posted on Twitter. This algorithm focuses on the original text of the messages posted on the network, discarding the other available metadata, except the timestamp that indicates the time the message was published.

[3]Facebook: http://www.facebook.com

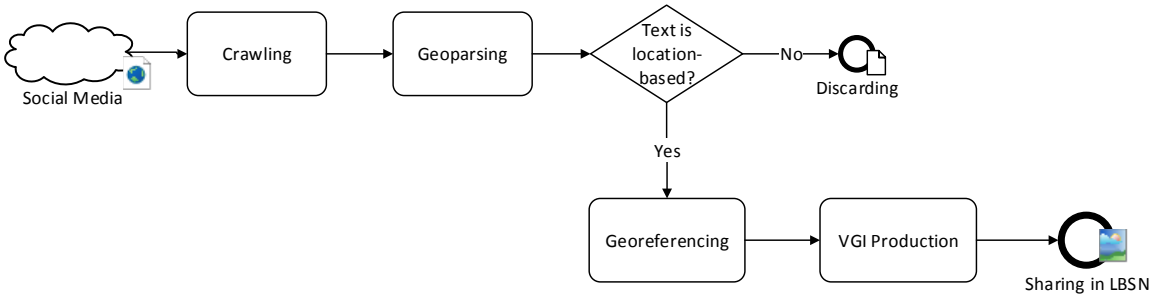

Fig. 2. Computational processing flow for automated VGI production

Once captured in the crawling stage, the microtexts are submitted to the geoparsing stage. In order to accomplish this stage, we used the GeoSEn Geoparser [Campelo and de S. Baptista 2009], which is capable of detecting geographic terms in texts written in Portuguese. At this stage, all candidate locations are identified and then sent to the next stage, in which the text will be georeferenced. Figure 3 illustrates the geoparsing and the georeferencing stages applied to a sample microtext.

In the upper half of Figure 3, it is possible to view all candidate locations identified in the sample microtext. The geoparser considers information such as the position of the term in the text and its length, that is, the number of words that form the term. The term position can be used to correlate spatial terms which may appear closely in the messages. In the case where the geoparsing of a microtext returns an empty set of candidate locations, this microtext is discarded and its VGI production process is interrupted. Geoparsing issues such as place names ambiguity and cross-references are treated by the GeoSEn Geoparser [Campelo and de S. Baptista 2009].

The georeferencing stage is illustrated in the lower half of Figure 3. In this stage, the candidate locations pass through a relevance evaluation in order to define the geographic scope of the microtext. We can notice that only one of the two candidate locations highlighted during the geoparsing stage was considered for the georeferencing of the sample microtext. Since one of these locations (the Dublin City) is inside the other one (Ireland), the geographic scope modeling algorithm returned just the most geographically precise one. For such, we used the GeoScope Modeler featured by the GeoSEn [Campelo and de S. Baptista 2008] in order to define the geographic scope and compute the relevance for its highest hierarchical levels, based on references found in lower levels. The GeoScope Modeler uses the GeoTree, a tree-based data structure which establishes the hierarchical relationship between places stored into the GeoSEn gazetteer. There are six hierarchical levels of places in the GeoTree: Country, Region, State, Mesoregion, Microregion and City. The first level represents regions with less geographic detail while the last represents regions with more geographic details. On the other hand, in case that more than one location are found and they do not have a parent relationship, all detected locations are returned by the geographic scope modeling algorithm. That implies the microtext will produce more than one VGI.

Finally, the VGI production stage is responsible for creating the spatiotemporal marker that will be shared on the LBSN. The marker is basically formed by the original microtext captured from the social network, the latitude and longitude coordinates obtained from the georeferencing stage and the timestamp of the moment that the message was first published in the social network. To generate spatial markers, we compute the centroid points of the geometries corresponding to the georeferenced texts. Moreover, these markers are produced automatically and assigned to an particular category defined in the Crowd4City so that they can be easily distinguished from the categories originally managed by the LBSN users, such as education, transportation and security. Thus, this exclusive category emphasizes that the spatial marker was not produced by a LBSN user.

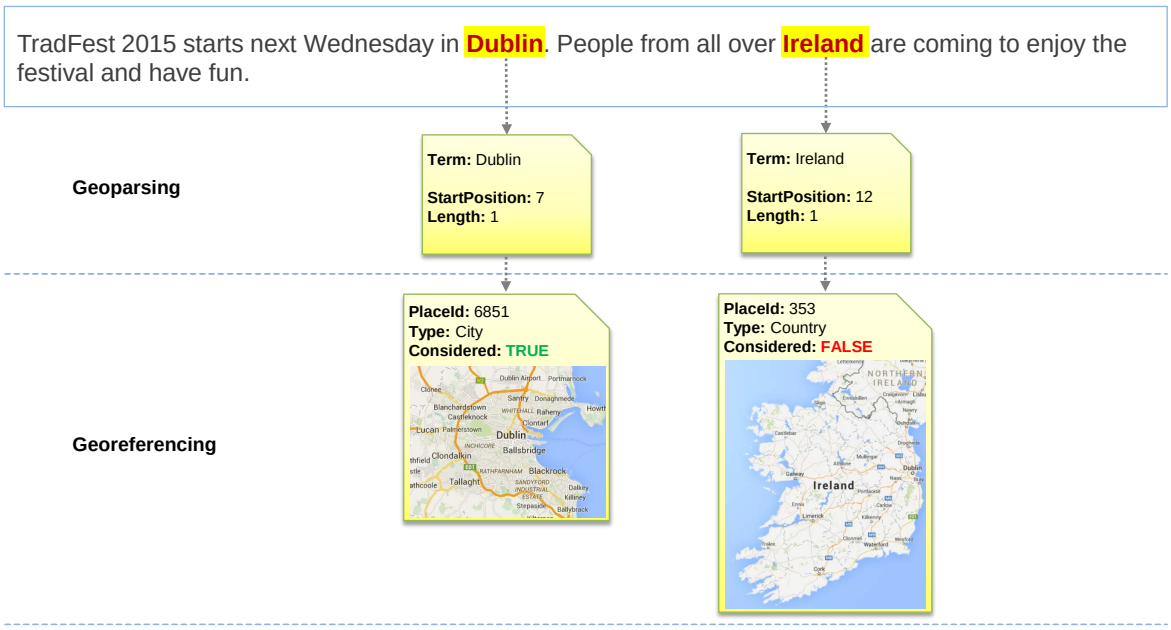

Fig. 3. Illustration of geoparsing and georeferencing stages applied to a sample microtext (*translated from Portuguese*)

A software application called *text2vgi* was implemented taking into account the whole flow illustrated in Figure 2. The purpose of such application is to validate the proposed approach, to confirm our hypothesis (discussed in the introductory section), and to identify points which may possibly require further improvements in order to ensure the most spatially-accurate VGI production.

3. CASE STUDY: PRODUCING VGI FROM NON-GEOCODED MICROTEXTS

This section presents a case study using the *text2vgi* software application with microtexts from Twitter.

3.1 The Dataset

Our case study was based on a dataset formed by 329,732 microtexts written in Portuguese, published on Twitter during the FIFA's Confederations Cup, which took place in Brazil in 2013. We adopted this dataset because it is related to an event in which people normally write terms that can be associated to geographic location, such as the name of the host cities. The methodology used for conducting this study is illustrated in Figure 4.

The Crawler implemented within the *text2vgi* tool is responsible for capturing the messages and storing them in a local database. As the messages are received by the application, the geoparser is activated to identify the candidate locations. Then, the georeferencing module defines the geographic scope of the microtexts that presented at least one candidate location. Finally, the VGI production module concludes the work by creating the spatiotemporal marker.

3.2 Manual Evaluation

The whole set of microtexts processed by *text2vgi* had to be evaluated in relation to the identified geographic locations and the spatiotemporal markers produced. By doing this, we could measure the quality of the VGI automatically produced. For such, seven volunteers were recruited and trained to use a web application we developed to help them conduct their analyses. This application displays

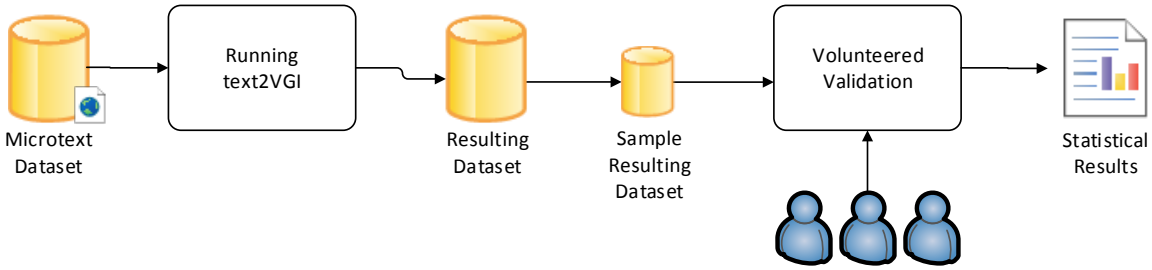

Fig. 4. The process flow for the case study

to the user a random list of processed microtexts, which should be individually analyzed by the volunteer. For each microtext, the user should evaluate the following characteristics: the overall geoparsing accuracy (boolean, checked-stars[1 to 5]); whether the message refers to more than one place (boolean); whether the location assigned to the message by the geoparser could be more precise (boolean).

The question regarding the geoprocessing accuracy could have as answer the pair (TRUE, 5-CHECKED-STARS) in the cases in which the georeferencing was totally accurate according to the georeferencing strategy used, or in the cases where the VGI was not produced because the microtext had no references to geographic locations. The answer to this question could also be the pair (FALSE, [0 | 1 | 2 | 3 | 4]-CHECKED-STARS), depending on the geographic and semantic distances between the location identified by the system (through geoparsing the microtext) and the actual location the microtext refers to (manually identified by the volunteers). An example of geographic distance is the case where a microtext refers to the city of "Campina Grande", whilst the system defines the microtext's geographic scope as "Paraíba" (the Brazilian State this city is located in) or "Nordeste" (the Region this state is located in). The semantic distance is often related to the system's misinterpretation of a term in the text, such as the case where the microtext's geographic scope is defined as "Bahia" (a Brazilian State), while it actually refers to the "Bahia" Football Team.

The question about whether a microtext refers to more than one place could have TRUE as answer if the microtext refers to more than one geographic location and, therefore, the system should produce more than one spatiotemporal marker for the same microtext; or FALSE, otherwise. Finally, the question about whether a microtext can be more precise could have TRUE as answer if the microtext presents evidences that its geographic scope could be more precise than the city level, such as neighborhood names, streets, squares, parks or specific buildings, such as stadiums.

From the whole set of processed microtexts, 2.3% (about 7,500) had at least one geographic location automatically assigned by *text2vgi* and, consequently, could be used to produce spatiotemporal markers. The manual evaluation helped us to understand that this low rate was not due to inefficiency of our geoparser. Oppositely, it could be observed that the geoparser efficiency was satisfactory, but indeed the majority of the messages actually do not refer to any geographic location.

Considering the huge volume of microtexts of the dataset used in this study, a random sample [Kish 2004] of these microtexts needed to be defined so that it could be evaluated by volunteers. Such a sample consisted of 35,120 microtexts, with a confidence level of 99% and a sampling error of 0.65%. In this sample, 975 microtexts (2.7%) had a geographic location automatically assigned by *text2vgi*, nearly the same proportion presented by the whole set of processed microtexts. Since the validation was performed by humans, we also considered a margin of error of 2.0%.

3.3 Discussion

The mean time for processing each microtext by *text2vgi* (from the moment of the capture of the message to the production of the spatiotemporal marker) was of 0.25 seconds. It took nearly 23 hours to process the whole dataset using a typical desktop computer, equipped with an Intel Core i7 processor, 8 GB of RAM and 1 single thread.

Considering the sample evaluated by the volunteers, Figure 5 presents the results for true positives, when the geographic location was identified correctly; false positives, when the geographic location was not identified correctly; true negatives, when there was no geographic scope assigned to the text, due to the lack of evidence in it; and false negatives, when no geographic scope was assigned to the text, but there was evidence for it.

Figure 5a shows that there was a balance between true and false positives, if we consider as true positives only the 100% precise location detections. In Figure 5b, it is possible to see the false positives in five classification levels. Each classification level represents how geographically close the false positive was to a true positive. It can be noticed that the false positives that are very far from the location expressed in the microtext (which received no stars in the accuracy question), represent only 24.6% - about half the total number of false positives. Finally, in Figure 5c, it is possible to observe a good result for true negatives. It confirms the lower rate of the processed microtexts which had at least one geographic location automatically assigned by *text2vgi*: in fact there were many microtexts that refer to no locations.

The evaluation performed by the volunteers on the microtexts also revealed in the following results:

—16.6 % of the microtexts presented evidences that more detailed geographic locations could have been identified. Thus a georeferencing strategy which can deal with a broader range of locations may improve the overall accuracy;

—3.2 % of the microtexts presented evidences that more than one geographic location could have been identified, producing therefore more than one spatiotemporal marker.

We have used four metrics for evaluating the overall performance of VGI produced automatically by *text2vgi* and validated by volunteers during this case study: Overall Accuracy (74.1%); Precision (92.3%); Recall (52.6%); and the F-Measure (0.67). F-Measure is the harmonic mean of precision and recall, widely used to evaluate classification algorithms in GIR [Martins et al. 2005]. Among such achieved values, it can be noticed a low recall rate, which means that 47.4% of the microtexts containing genuine references to geographic locations were not correctly interpreted by *text2vgi*.

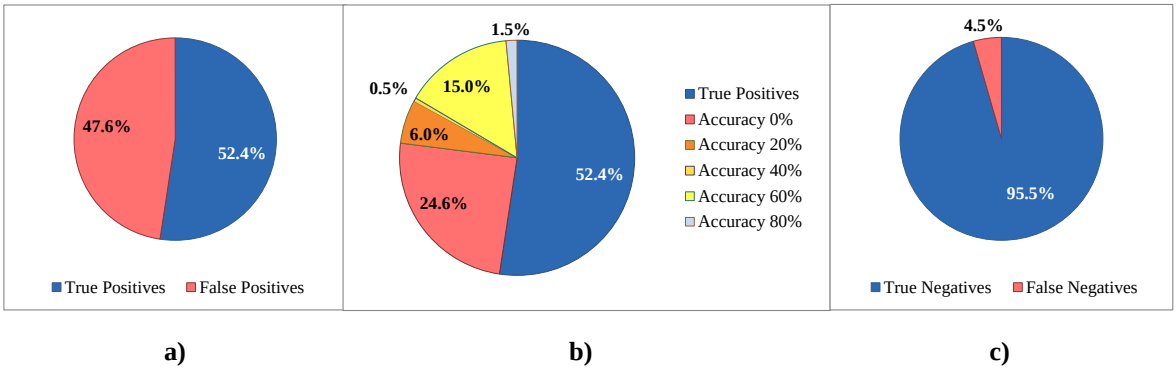

Fig. 5. Pie charts representing the percentages of each result: a) True/False Positives Relation, b) True/False Positives Relation considering the False Positives in five subdivisions, and c) True/False Negatives relation

However, this result was already expected, since the geographic scope considered in our georeferencing strategy considers only locations related to the Brazilian political territorial division. Other relevant geographic references that might appear in the set of microtexts used (such as soccer stadiums and airports) ended up not being properly interpreted. However, it is important to highlight the good precision rate, which is justified by the number of true negatives.

4. RELATED WORK

Research on VGI has prevailed in several parts of the world. Besides Computer Science, many correlated disciplines, such as Geography [Goodchild 2007], Geographic Information Science [Jackson et al. 2010] and Human Factors [Parker et al. 2011] have investigated issues concerning this kind of volunteered information.

One of the most representative VGI project is the OpenStreetMap (OSM) [Haklay and Weber 2008]. The OSM database consists of a significant collection of volunteered spatial data based on the Wikipedia collaborative model [Mooney and Corcoran 2012]. The OSM project has received many contributions from the community. Haklay [2010], for instance, has focused on assessing VGI quality and how VGI can be reliable and usable. Ballatore and Bertolotto [2011] focused on semantic relationships within OSM data. They highlight how OSM is spatially rich but semantically poor and investigate ways of linking OSM to other distributed repositories.

Besides the OSM project, several works have revealed VGI as a promising research field. Horita et al. [2013] made a thorough literature review on VGI with the objective of verifying its applicability for aiding in disaster management. In that study, it was possible to observe that the VGI has been more frequently used in fires and floods. Havlik et al. [2013] discussed VGI mobile applications concerning several aspects, such as functionalities and user experience. Ballatore et al. [2013] explored the semantic side of VGI and presented a technique for computing the semantic similarity of geographic terms in VGI based on their lexical definitions and using WordNet. The authors based themselves on the intuition that similar terms tend to be recursively defined by similar terms.

While the research on VGI is still relatively novel, the research on GIR has many studies focused on the identification and indexing of geographic locations through the application of Natural Language Processing (NLP) techniques. Rupp et al. [2013] discussed the customization of geoparsing and georeferencing tools to be applied in collections of historical texts. The authors made an analogy between the storage/indexing of files about the medieval era and the storage/indexing of Twitter feeds, and discussed questions involving standardization and use of gazetteers. There is no discussion about the spatial precision of the geoparsing, but this could be a motivational factor for such customization.

Liu et al. [2013] proposed the QGIR, Qualitative Geographic Information Retrieval, as a better option to deal with geographic information described in natural language in web documents. The authors argue the replacement of GIR by QGIR for cases where the place name and thematic representations are necessary, considering the use of semantic spatial relations and domain-specific ontologies. Freire et al. [2011] described an approach for recognizing place names expressed in metadata of digital libraries. That approach should be better at capturing features of the non-structured text found in metadata records and at the exploration of the relevant information in the structured data of those records.

Lee et al. [2013] leveraged the amount of geotagged tweets available and performed analysis in order to discover behavioural patterns. They only apply geoparsing when the location metadata is in a raw text format. Such work keep the focus on users' location and time without deeply analyzing their textual messages in order to identify spatial areas eventually mentioned. Following the same direction, Magdy et al. [2014] present a system for scalable querying, analyzing and visualizing geotagged microblogs which only considers pre-geotagged tweets. Neither geoparsing nor geotagging techniques are discussed or even applied by them. Hawelka et al. [2014] focused on human mobility through analyzing

geotagged Twitter messages containing explicit geographic coordinates. They could show the power of geotagged tweets, even with a tiny percentage of the messages, discovering human behaviours based on both space and time.

Watanabe et al. [2011] proposed an automatic method of identifying geographic location in non-geotagged tweets. Such method is based on the clustering of messages according to the type of event, considering short time intervals, small geographic areas and geotagged tweets. Thus, geotagged tweets are used to allocate geotags in tweets which do not have the geographic tag yet. The authors did not consider the possibility of the geotagged tweets having a different geographic reference than the location discussed in the messages. In addition, users do not necessarily talk about their current locations. Therefore, there is a possibility of errors in the geographic precision and this must be considered. In a similar way, Jung [2011] presented a method of analyzing sets of microtexts, aiming to identify contextual clusters of tweets. By establishing a contextual relation between the messages, a set of microtexts can be considered as a single document and make the process easier for the geoparsers. This task, however, can be very costly, depending on the volume of related tweets. In addition, there is also a possibility of errors in the geographic precision.

As we can notice, there are many researches addressing the power of location-related social media. While most authors focus on the user location through pre-geotagged metadata, others investigate the identification of geographic locations in social media messages focusing on exclusively on the text. However, the majority do not address specific issues of the Portuguese language. Furthermore, they do not address the domain of LBSNs and the aim of providing valuable information for such environments in an automated way. In this sense, our proposal comes as a solution to cover this gap.

5. CONCLUSION AND FURTHER WORK

In this article, we presented an approach to the automated production of VGI based on geoparsing and georeferencing of texts published on the Web. Such approach was conceived with the objective of turning Web authors into volunteers in the VGI context, contributing to the indirect production of information in a Location-based Social Network. A prototype, called *text2vgi*, was implemented with the goal of validating the ideas proposed by our approach. In order to evaluate the prototype in a real context, we carried out a case study using a set of microtexts in the Portuguese Language concerning a sporting event of large impact on the media, the 2013 FIFA's Confederations Cup, held in Brazil.

Overall, the achieved results were considered satisfactory. However, we have confirmed the need for improving the georeferencing strategy in order to increase the amount of VGI produced from microtexts, to improve the spatial accuracy of the spatiotemporal markers created and to achieve better results for the recall and F-Measure. It is important to consider points of interest (such as soccer stadiums and airports) and well known places in a city context. Thus, the automatically produced VGI will become more spatially precise and the user's experience in the LBSN will be improved.

As future work, we consider the implementation of georeferencing strategies to address the specific treatment of microtexts, such as the analysis of informal language. Furthermore, we will seek the development of heuristics that increase the precision of the locations detected, and consequently improve the F-Measure. Other future direction of our work is to improve our approach for producing VGI based on microtexts in other languages such as English, Spanish and French.

REFERENCES

Ballatore, A. and Bertolotto, M. Semantically Enriching VGI in Support of Implicit Feedback Analysis. In *Proceedings of the 10th International Symposium Web and Wireless Geographical Information Systems*. Kyoto, Japan, pp. 78–93, 2011.

Ballatore, A., Wilson, D. C., and Bertolotto, M. Computing the Semantic Similarity of Geographic Terms Using Volunteered Lexical Definitions. *International Journal of Geographical Information Science* 27 (10): 2099–2118, 2013.

Bordogna, G. and andGiuseppe Psaila, G. G. Geographic Information Retrieval: modeling uncertainty of user's context. *Fuzzy Sets and Systems* vol. 196, pp. 105–124, 2012.

Campelo, C. E. C. and de S. Baptista, C. Geographic Scope Modeling for Web Documents. In *Proceedings of the 2nd International Workshop on Geographic Information Retrieval*. NY, USA, pp. 11–18, 2008.

Campelo, C. E. C. and de S. Baptista, C. A Model for Geographic Knowledge Extraction on Web Documents. In *Proceedings of the Workshop on Advances in Conceptual Modeling - Challenging Perspectives*. Gramado, Brazil, pp. 317–326, 2009.

Falcão, A. G. R., de S. Baptista, C., and de Menezes, L. C. Crowd4City: utilizando sensores humanos como fonte de dados em cidades inteligentes. In *Proceedings of the 8th Brazilian Symposium on Information Systems*. São Paulo, Brazil, pp. 144–149, 2012.

Freire, N., Borbinha, J., Calado, P., and Martins, B. A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records. In *Proceedings of the 11th Annual International ACM/IEEE Joint Conference on Digital Libraries*. New York, NY, USA, pp. 339–348, 2011.

Goodchild, M. F. Citizens as voluntary sensors: spatial data infrastructure in the world of Web 2.0. *International Journal of Spatial Data Infrastructures Research* 2 (1): 24–32, 2007.

Haklay, M. How Good is Volunteered Geographical Information? A Comparative Study of OpenStreetMap and Ordnance Survey datasets. *Environment and Planning B: Planning and Design* 37 (4): 682–703, 2010.

Haklay, M. and Weber, P. OpenStreetMap: user-generated street Maps. *IEEE Pervasive Computing* 7 (4): 12–18, 2008.

Havlik, D., Soriano, J., Granell, C., Middleton, S. E., van der Schaaf, H., Berre, A. J., and Pielorz, J. Future Internet Enablers for VGI Applications. In *Proceedings of the International Conference on Environmental Informatics for Environmental Protection, Sustainable Development and Risk Management (EnviroInfo)*. Hamburg, Germany, pp. 622–630, 2013.

Hawelka, B., Sitko, I., Beinat, E., Sobolevsky, S., Kazakopoulos, P., and Ratti, C. Geo-located twitter as proxy for global mobility patterns. *Cartography and Geographic Information Science* 41 (3): 260–271, 2014.

Horita, F. E., Degrossi, L., de Assis, L. F., Zipf, A., and de Albuquerque, J. P. The use of Volunteered Geographic Information (VGI) and Crowdsourcing in Disaster Management: a systematic literature review. In *Proceedings of the 19th Americas Conference on Information Systems*. Chicago, Illinois, pp. 1–10, 2013.

Jackson, M. J., Rahemtulla, H., and Morley, J. The Synergistic Use of Authenticated and Crowd-sourced Data for Emergency Response. In *Proceedings of the 2nd International Workshop on Validation of Geo-Information Products for Crisis Management (VAL-gEO)*. Ispra, Italy, 2010.

Jung, J. J. Towards Named Entity Recognition Method for Microtexts in Online Social Networks: a case study of Twitter. In *Proceedings of the International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. Kaohsiung, Taiwan, pp. 563–564, 2011.

Kish, L. *Statistical Design for Research*. Wiley, Hoboken, New Jersey, 2004.

Lee, R., Wakamiya, S., and Sumiya, K. Urban Area Characterization Based on Crowd Behavioral Lifelogs over Twitter. *Personal and Ubiquitous Computing* 17 (4): 605–620, 2013.

Liu, L., Gao, Y., Lin, X., Guo, X., and Li, H. A Framework and Implementation for Qualitative Geographic Information Retrieval. In *Proceedings of the 21st International Conference on Geoinformatics (GEOINFORMATICS)*. Kaifeng, China, pp. 1–4, 2013.

Magdy, A., Alarabi, L., Al-Harthi, S., Musleh, M., Ghanem, T. M., Ghani, S., and Mokbel, M. F. Taghreed: a system for querying, analyzing, and visualizing geotagged microblogs. In *Proceedings of the 22nd ACM SIGSPATIAL*. Dallas, Texas, USA, pp. 163–172, 2014.

Martins, B., Silva, M. J., and Chaves, M. S. Challenges and Resources for Evaluating Geographical IR. In *Proceedings of the 2005 Workshop On Geographic Information Retrieval (GIR 2005)*. Bremen, Germany, pp. 65–69, 2005.

Mooney, P. and Corcoran, P. The Annotation Process in OpenStreetMap. *Transactions in GIS* 16 (4): 561–579, 2012.

Parker, C. J., May, A., and Mitchell, V. Relevance of Volunteered Geographic Information in a Real World Context. In *Proceedings of the GIS Research UK Conference (GISRUK)*. Portsmouth, UK, pp. 230–236, 2011.

Purves, R. and Jones, C. Geographic Information Retrieval. *SIGSPATIAL Special* 3 (2): 2–4, 2011.

Rupp, C. J., Rayson, P., Baron, A., Donaldson, C., Gregory, I., Hardie, A., and Murrieta-Flores, P. Customising Geoparsing and Georeferencing for Historical Texts. In *Proceedings of the IEEE International Conference on Big Data*. Silicon Valley, CA, pp. 59–62, 2013.

Surowiecki, J. *The Wisdom of Crowds*. Anchor, New York, USA, 2005.

Vicente, C. R., Freni, D., Bettini, C., and Jensen, C. S. Location-Related Privacy in Geo-Social Networks. *IEEE Internet Computing* 15 (3): 20–27, 2011.

Watanabe, K., Ochi, M., Okabe, M., and Onai, R. Jasmine: a real-time local-event detection system based on geolocation information propagated to microblogs. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*. New York, NY, USA, pp. 2541–2544, 2011.