Building Social Machines from Social Networks Data

Eduardo Santos
Departamento de Ciência da Computação
Universidade de Brasília
Campus Universitário Darcy Ribeiro
Brasília/DF – Brasil, CEP 70910-900
Email: eduardo@gnumasters.com

Fernanda Lima
Departamento de Ciência da Computação
Universidade de Brasília
Campus Universitário Darcy Ribeiro
Brasília/DF – Brasil, CEP 70910-900
Email: ferlima@cic.unb.br

***Abstract*—Social machine is a rather new approach to deal with relevant problems in society, blending computational and social elements into software. It can be an extension of the Semantic Web, creating processes in which people do the creative work and the machine does the administration. This article presents a proposal to apply this approach in a relevant matter to Latin America and Caribbean (LAC) countries. The result is demonstrated by the implementation of a theoretical solution in violence and criminality domain.**

I. Introduction

In June 2013, when FIFA Confederations Cups started, Brazil saw a general uprising explode in an wave of protests that began in south region and spread almost everywhere in the country. The biggest impact fact is that none of country's ruling powers and institutions was able to predict it was coming, as "social media boost a protest's transmission rate through susceptible societies" [1]. This example supposes that social networks are free environments for people to share their frustrations. Talking to one million people can be as far as a mouse click, and Brazilian protests page [2] is a good example [1]. The surprise all government institutions demonstrated at the time suggests most of this dissatisfaction keeps off the track for official records.

The concept of social machines, web technologies to deal with real society problems [3], can be a way to minimize the information gap between government policies and citizens needs. "Unstructured data, unreliable parts and problematic, non-scalable protocols are all native characteristics of the Internet that has been evolving for 40 years" [4]. It is necessary to think about Web as a platform of connected services, and social networks represent one important node on this network graph. "Web-based social software (collectively called 'Web 2.0' which consists of blogs, social networking websites, video sharing, etc) can be seen as early versions of Social Machines" [5].

Such interaction technology supposes the introduction of human activity as a computational tool. "Does a social machine have to incorporate a 'machine' in the sense that we might think of a computer, or can machine be used in the wider sense, as in some sense of a Turing Machine; a series of computations" [6]? This is one of the main definitions for crowdsourcing, and human interaction efficiency can be improved if the problem is relevant enough so more people are willing to participate [7].

According to recent data, violence is the second biggest problem for 18% of the population in Brazil. It comes right after health, which leads for 45% of the population [8]. The insecurity perception is also part of citizens' lives in Brazil and other Latin America and Caribbean (LAC) countries: it is the region with the highest murder rate in the world. The UN study about cities in LAC explains this issue through the social inequality [9][p.XII]: as there are more people living in good and wealthy conditions, there even more living in total poverty. The population lives in a social tension atmosphere, which leads to violence in the end.

Even though it is an important issue for the population, only in the last 10 years Brazilian government created an unified system to centralize criminal data. Before 2003, with the creation of Unified Public Security System (*Sistema Único de Segurança Pública* – SUSP in portuguese), "the management of policy and security actions where characterized by the absence of cooperation between organizations" [10]. The statistics relied on manual delivery from the police stations to federal government. Even in more developed countries as UK, the criminal data represent a real problem: "from deciding a crime has occurred, to reporting and recording, there are areas in which the data can mislead" [6]. The authors propose and analyze a social machine solution based on crowdsourcing Open Crime Data.

The theory and practice of social machines is currently being addressed by several authors with different goals. In theory definition, formal models lack validation on more specific domains. "The science, technology and implementations of Social Machines are in a very early stage. (...) Future work includes extending this study to tackle an even more comprehensive set of references and the generalization of the results to include other aspects of the research area" [5].

This work presents the concept of a social machine as a crowdsourcing system to identify violence and criminality on social networks activity. It works on the gap of generalizing the results from systematic mapping studies [5] on violence and criminality domain.

[1] A facebook page connected to protests in Brazil went from thousands of members to hundreds of thousands overnight

II. Social Machines

Knowledge can be defined as "the accumulation of techniques for addressing the specific aspirations of each era", and these techniques are defined as technologies [11]. The knowledge acquisition technologies addressed by Gaines are part of his definition about Symbiosis Science, starting with Web 2.0 phenomena [12]. Hendler and Berners-Lee advocate that "just as human communities interlink in society they must be interlinked on the Web" [13].

Analyzing the huge amount of information available on the Web, it is necessary to review the way people interact to create applications at a higher level of abstraction than just tables presented on a website. The new model would behave like a "global graph of interconnected people and ideas" [13], and it should be able to blend social and computational processes.

A. From the Semantic Web to Social Machines

The foundations for social and computational connection technologies lie on the initial vision of the Semantic Web [14]. The proposal was to extend Web to give a well-defined meaning for information, "better enabling computers and people to work in cooperation". With semantic added to Web documents, it would be possible to create a global network of documents without concerning about how they would be stored and exchanged.

The challenge of creating typed links between different networks is defined as Linked Data [15]. The growth of linked open data applications development was represented by the Linked Open Data Cloud Project. The cloud is an abstraction from the Linked Open Data Project [15].

The described scenario generates a problem with "the segmentation of data and the issues regarding the communication among systems" [4]. Even though the Open Data movement started with the main motivation of liberating information, "none of the resources are the slightest good without context and schemes that enable one to interpret what they are about" [3]. Some mapping between computer processed data and society needs is necessary, suggesting that to interpret linked data a new kind of computation with social context arises.

B. Social Computation

The example from FIFA Conferations Cup presented in section I talks about a leaderless network away from hierarchy based on self-organization. In fact, even though the protests suggests a lack of political organization, it doesn't mean they are free from political conscience. A study analyzing geographical locations of protests participants in twitter [2] proves, among other hypothesis, that users focused their online participation on wealthier regions. This data can show some consciousness about how important it was to aim on political power structures in their online protesting activity.

The example suggests that some kind of human intervention to address society needs must be added to social networks data interchange, even more if we consider mash ups with linked data. Shadbolt defines the core of social machines at many stages [3], situating the definition as a knowledge acquisition system:

Crowsourcing
: Users are able to define relevant information by themselves and share their personal rankings with each others.

Social computation
: The collaboration is somehow introduced back to the social application and helps other users.

C. Abstract Model for Social Machines

Crowdsourcing is usually the first step of social computation. To build social machines according to the aforementioned first definition, one can create applications capable of introducing relevant information as a service to help other users solve their social demands. Other authors contributed to the notion of Social Machines [5] performing a systematic mapping study. The authors organized the definitions in views and and introduced a very important aspect: software as sociable entities. With the purpose of defining a formal algebra around three views, they proposed the following models:

People
: Defined as computational unit, the vision of people involve tasks and behaviors possible only to humans, such as *crowdsourcing*;

Social software
: Technologies for the web that involve some kind of information exchange between users, such as social networks in Web 2.0 and API platforms like twitter and facebook;

Software association
: Software and services for the web that contain one or more integration layers with other social software like themselves, such as linked data applications or social networks with other web-services.

The three views allow us to think about social machines as an "abstract model to describe Web based information systems that could be a practical way of dealing with the complexity of the emerging programmable web" [16].

A formal definition is presented by [4] as the equation 1.

$$SM = < Rel, WI, Req, Resp, S, Const, I, P, O > \quad (1)$$

According to the authors, it represents the mental model for the web as a platform, describing relations between connected services. One attempt to validate the definition is presented by the *WhatHere* application [16], implementing one that works as social machines composition. One extension of the previous work is presented by the Social Machines Architecture Languages definition – SMADL –, a domain description language to describe social machines [17].

III. Criminal data

When proposing a framework for crime data mining, [18] presents entity extraction technique:

> Entity extraction identifies particular patterns from data such as text, images, or audio materials.

> It has been used to automatically identify persons, addresses, vehicles, and personal characteristics from police narrative reports

The author presents an analysis between crime data mining techniques and the capability of identification concerning the crime classification. According to the article, entity extraction is the only one capable of identifying with acceptable accuracy all types of crimes.

Definition of criminal activity is the first step in analyzing and understanding violence data. As stated in [18], "a criminal act can encompass a wide range of activities, from civil infractions such as illegal parking to internationally organized mass murder such as the 9/11 attacks". Police reports in Brazil use *incident* concept, the occurrence of a criminal act of any kind [19][p.148].

A. Criminal Databases

In USA, FBI keeps National Incident-Based Reporting System [20] – NIBRS – to organize the incident-based criminal data in two categories: Group A, for which extensive crime data are collected and Group B, for which only arrest data are reported. This system is part of Uniform Crime Record – UCR (http://www2.fbi.gov/ucr/ucr.htm) project, which is responsible to generate uniform crime statistics for the country. NIBRS data comes from more than 5,000 agencies representing 20% of the population and 16% of all crime statistics.

In Brazil, it is still a challenge to build an uniform database of criminal incidents. The design of an Unified Public Security System (*SUSP* in Portuguese), presented in [19][p.141], supposes a three level data gathering. The data flow order is presented at Figure 1. The authors defend that the information should be gathered as close to the source as possible, starting at the municipality.

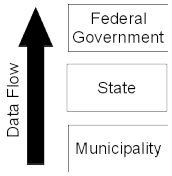

Figure 1. Data flow in criminal data

B. Criminal reports

Brazilian law also defines different police types for different activities: military police is responsible for crime prevention, while civil police answers for crime investigation. Both of them use different systems and maintain different databases. The incident registration process is described in Figure 2. From the incident to the official report, the first step is going through military or civil police. Even if the incident is registered by the military police, it's still necessary to pass the information to the civil police, where the incident will have three different destinations:

1) "Balcony closure" or police mediation for civil conflicts. When it happens, the police officer can try to solve the issue between the involved parts, without initiating a crime report;
2) Case forward to other institutions, in case the incident is not qualified as a crime and it doesn't require a police investigation;
3) Fulfill the police report. On this case the commissioner in charge can submit the incident to an administrative routine or to start a criminal investigation.

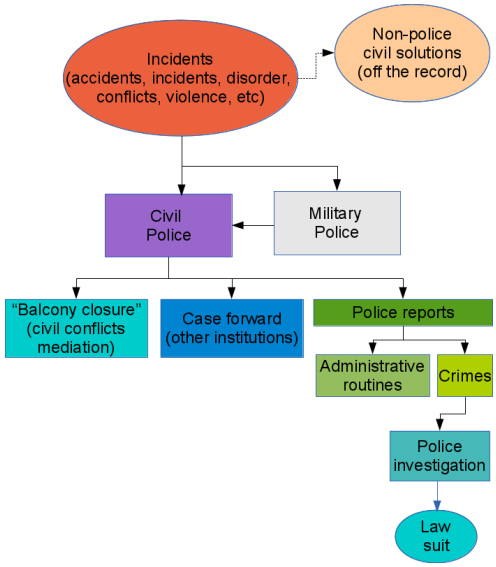

Figure 2. Crime reporting in Brazilian police [21]

The scenario complexity supposes it is difficult to register a police report. In fact, when thinking about the Unified Security System [21], the authors emphasize the problem:

> Most incidents (...) with the population are followed by non-police civil solutions as response. (...) This decision keeps a lot of incidents off the record to the public security institutions.

To bring data about this non-police civil information can help government and population to have a better understanding about violence and criminality.

IV. Theoretical Solution

The theoretical solution has two sides: an implementation proposal focused more on citizens than governments, and documenting the implementation for the definition of a systematic procedure. The first will focus on delivering the solution for population, while the second situates development work on the theory.

Without a full international government support, such application can not replace official tools. Instead of defining self regulation mechanisms to make sure the information is reliable, it is up to the users to *crowdsource* the data as they want. Instead of creating a formal environment to deal with all the social constraints, the proposal will build a peer to peer model to help users with non-police civil responses [21].

To build a theoretical solution the design guidelines from Evidence Based Software Engineering – EBSE – will be used [22]. The importance of population addressed in guidelines *D1* and *D2* was also addressed by [23] considering social

networks data. After using different aspects of Twitter API to collect information from users timeline, they chose to use a local news agency twitter as data source.

However, social machines should answer social demands, not only represent a filtered copy of news agencies agenda. The envisioned architecture should express a hybrid approach, making it clear to users where the data can be more reliable, but also giving them the opportunity to evaluate other sources they believe. It is the human computation aspect of social machines.

The final observation about design concepts comes from defining the experimental unit (*D5*). Official data about violence and criminality, as stated in section III, work with *incident* concept. SRL technique described in section II-A identifies *events* from tweets. The set of events identified in a topic structure represent the experimental unit in a crime and violence classification. Table I summarizes guidelines adopted and implementation proposal.

Guideline	Theory	Implementation
D1	Population	Twitter and facebook users
D2	Selection process	Reliable sources, crowdsourcing and users evaluation
D3	Extraction strategy	Semantic analysis of large-scale social networks data
D5	Experimental unit	Criminal and violence related incidents
D6	Precalculation	Twitter extraction accuracy analysis

Table I. RESEARCH GUIDELINES ADOPTED AS [22]

V. FINAL CONSIDERATIONS

The first words about social machines came in 2000 [24], but an important milestone was the introduction of social networks and linked data [13]. A conceptual model is being built [16], together with mathematical [4] and computational [17]. This effort goal is summarized by [11]: how can these technologies help in our everyday life?

This issue can be addressed in two different views: regular citizens, producing and consuming social data, and governments, elaborating public policies and regulations to address citizens needs. Social networks supply a general view on user surroundings, reading twitter timeline or sharing network activity through facebook. However, every communication channel has a limit about the number of connected nodes the message is able to achieve. Building an integrated technology is about finding hidden network nodes, which could not be achieved through regular channels.

Different from other examples showing processed data, this proposal should be able to show tendencies in almost real time. The expected output will contain biggest challenges during system development, technological constraints imposed by the architecture, theoretical background requirements and system usage manual. Produced source code will be published in a Free Software repository, together with produced documentation. Considering the gap of empirical validation, this study is part of the ongoing process of defining a theoretical and practical background in social machines.

REFERENCES

[1] D. MacKenzie, "Brazil's uprising points to rise of leaderless networks," *New Scientist*, vol. 218, no. 2923, p. 9, 2013.

[2] M. Bastos, R. Recuero, and G. Zago, "Taking tweets to the streets: A spatial analysis of the vinegar protests in brazil," *First Monday*, vol. 19, no. 3, 2014. [Online]. Available: http://firstmonday.org/ojs/index.php/fm/article/view/5227

[3] N. Shadbolt, "Knowledge acquisition and the rise of social machines," *IJHCS*, vol. 71, pp. 200–205, 2013.

[4] S. R. Meira, V. A. Buregio, L. M. Nascimento, E. Figueiredo, M. Neto, B. Encarnação, and V. C. Garcia, "The emerging web of social machines," in *COMPSAC, 2011 IEEE 35th Annual*.

[5] V. Burégio, S. Meira, and N. Rosa, "Social machines: a unified paradigm to describe social web-oriented systems," in *WWW Companion '13*, 2013, pp. 885–890.

[6] M. B. Evans, K. O'Hara, T. Tiropanis, and C. Webber, "Crime applications and social machines: crowdsourcing sensitive data," in *SOCIAM: The Theory and Practice of Social Machines*, May 2013. [Online]. Available: http://eprints.soton.ac.uk/351275/

[7] A. Doan, R. Ramakrishnan, and A. Y. Halevy, "Crowdsourcing systems on the world-wide web," *Commun. ACM*, vol. 54, no. 4, pp. 86–96, Apr. 2011. [Online]. Available: http://doi.acm.org/10.1145/1924421.1924442

[8] M. Leite, "Datafolha aponta saúde como principal problema dos brasileiros," 2014. [Online]. Available: http://folha.com/no1432478

[9] ONU-Habitat, *Estado de las ciudades de América Latina y el Caribe 2012*, 2012.

[10] M. O. Durante, "Avanços e desafios na implantação do sistema nacional de estatísticas de segurança pública e justiça criminal (SINESPJC)," *Anuário de Segurança Pública*, 2008.

[11] B. R. Gaines, "Knowledge acquisition: past, present and future," *IJHCS*, vol. 71, pp. 135–156, 2013.

[12] T. O'reilly, "What is web 2.0: Design patterns and business models for the next generation of software," *Communications & strategies*, vol. 65, no. 1, pp. 17–37, 2007.

[13] J. Hendler and T. Berners-Lee, "From the semantic web to social machines: A research challenge for ai on the world wide web," *ARTINT*, vol. 174, pp. 156–161, 2010.

[14] T. Berners-Lee, J. Hendler, O. Lassila *et al.*, "The semantic web," *Scientific American*, 2001. [Online]. Available: http://www.scientificamerican.com/article/the-semantic-web/

[15] C. Bizer, T. Heath, and T. Berners-Lee, "Linked data-the story so far," *IJSWIS*, vol. 5, no. 3, pp. 1–22, 2009.

[16] K. dos Santos Brito, L. E. A. Otero, P. F. Muniz, L. M. Nascimento, V. A. de Arruda Burégio, V. C. Garcia, and S. R. de Lemos Meira, "Implementing web applications as social machines composition: A case study." in *SEKE*, 2012, pp. 311–314.

[17] L. M. d. Nascimento, V. A. Burégio, V. C. Garcia, and S. R. Meira, "A new architecture description language for social machines," in *WWW Companion '14*, ser. WWW Companion '14, 2014, pp. 873–874. [Online]. Available: http://dx.doi.org/10.1145/2567948.2578831

[18] H. Chen, W. Chung, J. J. Xu, G. Wang, Y. Qin, and M. Chau, "Crime data mining: a general framework and some examples," *Computer*, vol. 37, no. 4, pp. 50–56, 2004.

[19] J. LEMGRUBER *et al.*, "Arquitetura institucional do sistema único de segurança pública," *Acordo de Cooperação Técnica: MJUS, FIERJ, SESI, PNUD*, 2004.

[20] FBI, "National incident-based reporting system (NIBRS) – General Information," 2014. [Online]. Available: http://www2.fbi.gov/ucr/faqs.htm

[21] BRASIL, "Crime records map," 2009. [Online]. Available: http://migre.me/kHPO5

[22] B. A. Kitchenham, S. L. Pfleeger, L. M. Pickard, P. W. Jones, D. C. Hoaglin, K. El Emam, and J. Rosenberg, "Preliminary guidelines for empirical research in software engineering," *Software Engineering, IEEE Transactions on*, vol. 28, no. 8, pp. 721–734, 2002.

[23] X. Wang, M. S. Gerber, and D. E. Brown, "Automatic crime prediction using events extracted from twitter posts," in *Social Computing, Behavioral-Cultural Modeling and Prediction*. Springer, 2012, pp. 231–238.

[24] T. Berners-Lee, M. Fischetti, and M. L. Foreword By-Dertouzos, *Weaving the Web: The original design and ultimate destiny of the World Wide Web by its inventor*. HarperInformation, 2000.