View metadata, citation and similar papers at core.ac.uk
brought to you by CORE
provided by DSpace at Open Universiteit Nederland

U.P.B. Sci. Bull., Series C, Vol. 79, Iss. 2, 2017 ISSN 2286-3540

BLOGCRAWL: CUSTOMIZED CRAWLING OF ONLINE COMMUNITIES

Lucia Larise STAVARACHE[1], Mihaela BALINT[2], Mihai DASCALU[3], Stefan TRAUSAN-MATU[4], Nicolae NISTOR[5]

With half of the world already connected to the Internet, we are facing a growing amount of information available online, that is expected to increase exponentially in the following years. Educational environments are transitioning from closed structures to open, collaborative environments, using technology to build virtual classrooms. In this paper we present a customized crawler dedicated to alternative knowledge building environments used for potential community inquiry, that is unique in its power to combine data extraction and indexing capabilities that facilitate discourse-driven community network analysis integrated into the ReaderBench framework.

Keywords: Online Communities, Crawling, Timeline Evolution, Knowledge Extraction

1. Introduction

The motivation to develop a customized crawler emerged while studying virtual learning communities [1, 2]**,** which involve large volumes of data having as order of magnitude thousands of participants and tens of thousands of contributions spanning multiple years. The aim has been to design a framework that performs data crawling and spidering using a consistent and structured model to map information from massive open online courses (MOOCs) [3], computer-supported collaborative learning (CSCL) technologies (e.g., forums, chats) [4], or learning communities onto an aggregated output format. In data extraction, the focus fell on the ability to mitigate security policies that block data crawling

[1] PhD Student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: larise.stavarache@ro.ibm.com
[2] Teaching Assistant and PhD Student, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: mihaela.balint@cs.pub.ro
[3] Associate Prof., PhD, Dept. of Computer Science, University POLITEHNICA of Bucharest, Romania, e-mail: mihai.dascalu@cs.pub.ro
[4] Prof., Faculty of Automatic Control and Computers, University POLITEHNICA of Bucharest; Senior Researcher, Research Institute for Artificial Intelligence of the Romanian Academy; full member of the Academy of Romanian Scientists, e-mail: stefan.trausan@cs.pub.ro
[5] P.D., Ludwig-Maximilians-Universität München, Germany/Walden University, USA, e-mail: nic.nistor@lmu.de

threads, replicating the conditions over multiple runs, and the recognition and mapping to a uniform representation of various unstructured input data. There are multiple crawlers available on the Internet who fail to offer a standardized method to gather clean data and analyze virtual learning communities. Specific features of our own solution, BlogCrawl, are outlined below, in comparison to other crawlers.

While *Repository Based Software Engineering* (http://www.robotstxt.org/db/rbse.html), a NASA funded data spider, crawls and downloads raw Internet pages, BlogCrawl uses a virtual Document Object Model (DOM), described later on in detail, to clean, normalize, and format the data into a structure that preserves essential discourse information (such as the inter-animation structure of the original conversation). Moreover, specific connectors enable the crawler to interface with Wordpress, BlogSpot, Coursera, LinkedIn, Twitter, Facebook, MOOCs, etc.

WebCrawler (http://www.webcrawler.com/), the first full text Web search engine, developed in 1995 by America Online, initially used a database storage model, but nowadays only focuses on metasearch – aggregating the top results from Google Search and Yahoo! Search. In contrast, BlogCrawl's download and analysis of data rely on an own Uniform Resource Identifier (URI) discovery and spidering model.

Googlebot, developed in 1998 by Sergey Brin, is the web crawler currently used by Google Search. Googlebot was designed to operate at a very large scale, hence it focuses on indexing, ranking, and discovery of new content. In contrast, BlogCrawl was designed to extract a specific kind of data (of academic interest). Relevant content is indicated by human analysts, and the crawler's job is to map this content to a standardized representation, suitable for the analysis of learning communities. Thus, there is no common ground between the two crawlers, and between BlogCrawl and search bots (e.g., BingBot, ExaBot) in general, apart from the principles behind the URI discovery algorithm.

To sum up, what differentiates BlogCrawl from these and other crawlers available in the open market (e.g., Nutch, Aperture, Scrapy, GNU Wget, GRUB, PHP-Crawler, WebSPHINX, Jspider, HyperSpider, crawler4j) is: (a) aimed at analyzing content starting from a list of URLs provided by the user; in turn, the extracted information is subject to automated content analysis [5] and may be used to predict how likely the examined online communities will respond to newcomer inquiries [6, 7], (b) a rigorous procedure to clean, normalize, formalize, and standardize data, (c) integration with ReaderBench [8] that enables complex Natural Language Processing [9] and discourse analyses [10], and (d) visualization graphs generated directly from the extracted data that facilitate the timeline analysis of the discourse threads from the user-selected educational conversational environments.

The remainder of this paper is structured as follows. Section 2 introduces the architecture and underlying technology of BlogCrawl, with an emphasis on its individualizing features: uniform data representation, visualization options, and integration with ReaderBench. Section 3 elaborates on possible uses and describes the experiments in which BlogCrawl has been validated. Finally, section 4 discusses the benefits of the framework for academic collaborative environments, and draws directions for further study.

2. Architecture and capabilities

As depicted in Fig. 1, BlogCrawl comprises four main components: a set of source connectors (used to handle multiple data sources), a crawling engine (used for data extraction and processing), solutions for data storage, and a generator of output data formats, and is able to access data either as a Java archive (JAR), or through its Representational State Transfer (REST) API. The crawler is compatible with multiple Java servers: JBoss, Tomcat, and WebSphere Liberty. BlogCrawl adopts a multi-threaded approach using rewind input stream (RIS) processing in memory under a preset limit of 2MB per document, without supporting parallel processing of the same document. BlogCrawl offers persistence by integration with different databases via dedicated connectors, thus enabling the storage, indexing, and lookup in big data. For experiments on small to medium corpora, the FileSystem storage is also an option embedded in the configuration model.

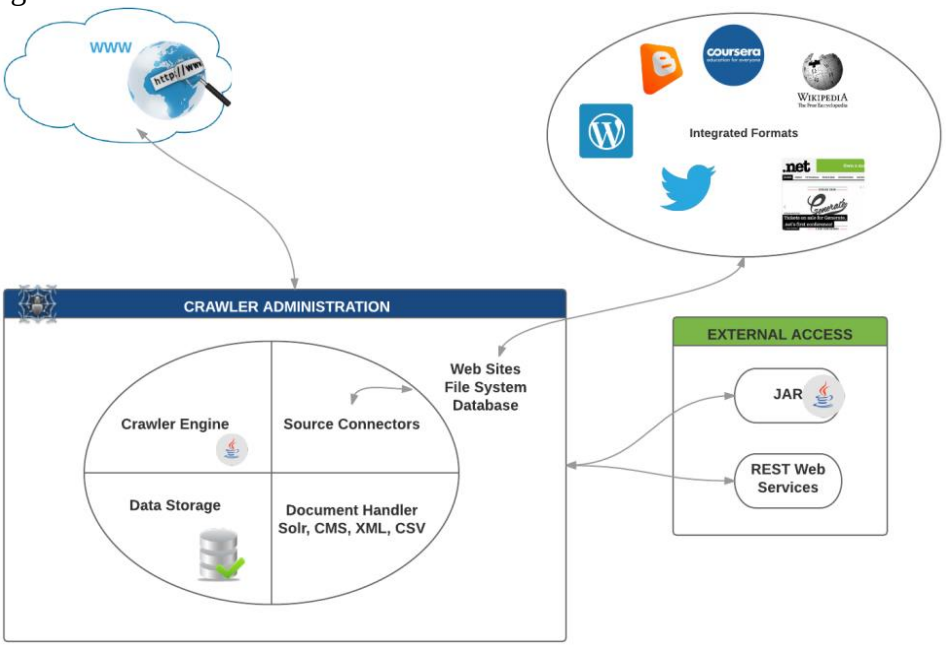


Fig. 1. BlogCrawl Logical architecture

Each source connector is designed to handle the specific challenges imposed by data sources such as Wordpress, Blogspot, Coursera, Wikipedia, Twitter, Net, Facebook, or straightforward custom MOOCs or discussions. For example, in Twitter, data extraction is easy (direct), while mapping the data to an informative structure needs to solve problems of language, or the usage of hashtag correspondence rather than end-to-end sentences. In contrast, Facebook carries content, but it has a policy that limits access to data. Furthermore, BlogCrawl works with Secure Sockets Layer (SSL; https://tools.ietf.org/html/rfc6101), which allows it to meet the security standards imposed by certain virtual environments.

Table 1

BlogCrawl's features

Features	Description
Programming language	Java 1.8 JDK
Build method	Maven
Interfaces	Jar, REST API
Integration sources	WordPress, Blogspot, .net, Wikipedia, Coursera, Twitter, Facebook
Supported sources	Any web page with conversational taxonomy
Data output model	Virtual HTML DOM
Database connectors	Oracle/DB2
File system	Supported
Multi-threading	Supported
Caching	Supported
SSL	Supported
Input formats	HTML, XML, CSV, txt, Excel
Output Formats	HTML, XML, CSV, txt, Excel, PDF

The crawling engine goes through several stages prior to data processing (e.g., URI discovery and filtering, DNS resolving, RIS - Rewind InputStream, link extraction, tag counting), followed by data collection, data cleaning, mapping data to BlogCrawl's virtual DOM format, and disposing of repetitive results. Solutions for data storage include both mechanisms for neighborhood stockpiling (as XML, XSV, Excel, HTML, TXT), and database stockpiling (Oracle, DB2). BlogCrawl can centralize its results in multiple output formats, including HTML, XML, CSV, TXT, Excel, and PDF. Table 2 offers a detailed overview of BlogCrawl's technical capabilities.

Uniform data representation

BlogCrawl is unique when compared to other crawling tools in that it maps multiple source formats to a uniform aggregated view (depicted in Fig. 2), following an XML/XSD validation output reflected in the virtual Document Object Model (DOM; https://www.w3.org/TR/DOM-Level-3-Core/). DOM is a language independent programming interface for XML, HTML, XHTML, and

other compatible formats, that represents the connections between elements or tags in a structured tree object. A virtual DOM extends the concept of DOM by introducing a custom structure, meaningful for data in terms of follow-up processing. In the virtual DOM representation, a community is represented from a structural perspective: participants, body of dialog, turns with corresponding utterances. The turns stand for members' interventions, and the discussion thread emerges as multiple participants share their views with respect to the main post or previous interventions. Each turn specifies a discourse participant and descriptive information for the associated comment: ID, timestamp, cross reference to the parent (0 if the comment responds to the main post), and actual text.

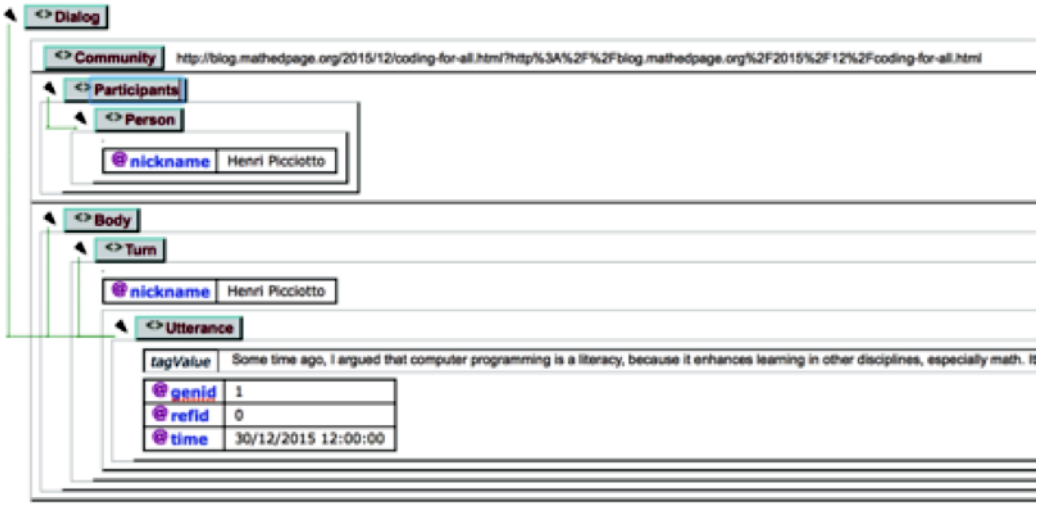


Fig. 2. Virtual DOM output structure.

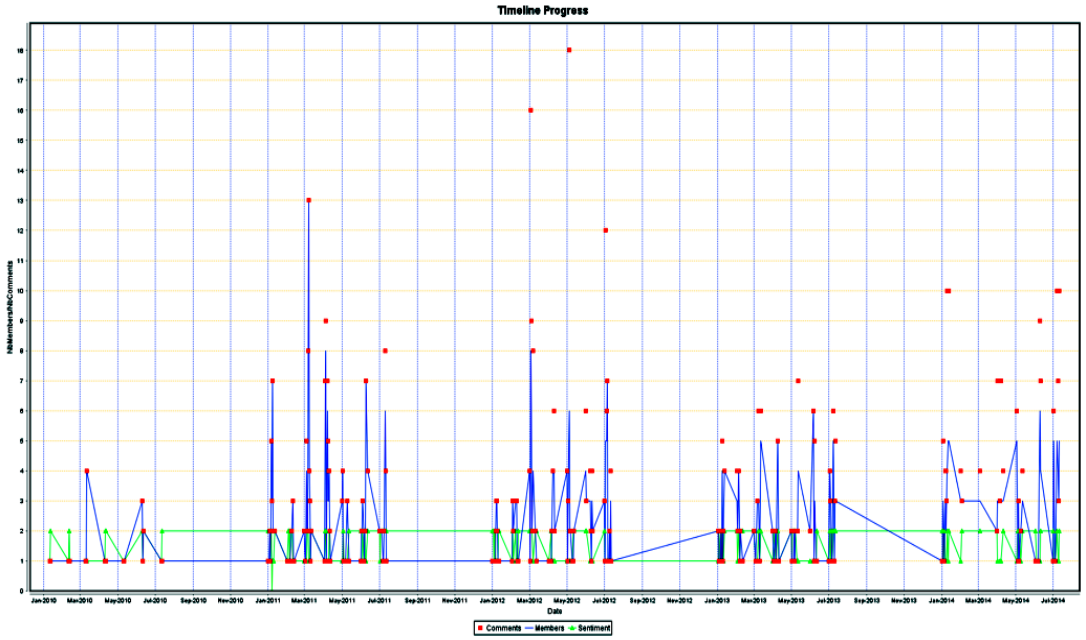
Visualization options

BlogCrawl includes a timeline evolution-modeling component for several community descriptors such as: number of members, number of posts and comments, sentiment or topic coverage associated with posts and comments, etc. The option to visualize the community's evolution in time from various perspectives provides valuable insights into the community's structural and collaboration patterns.

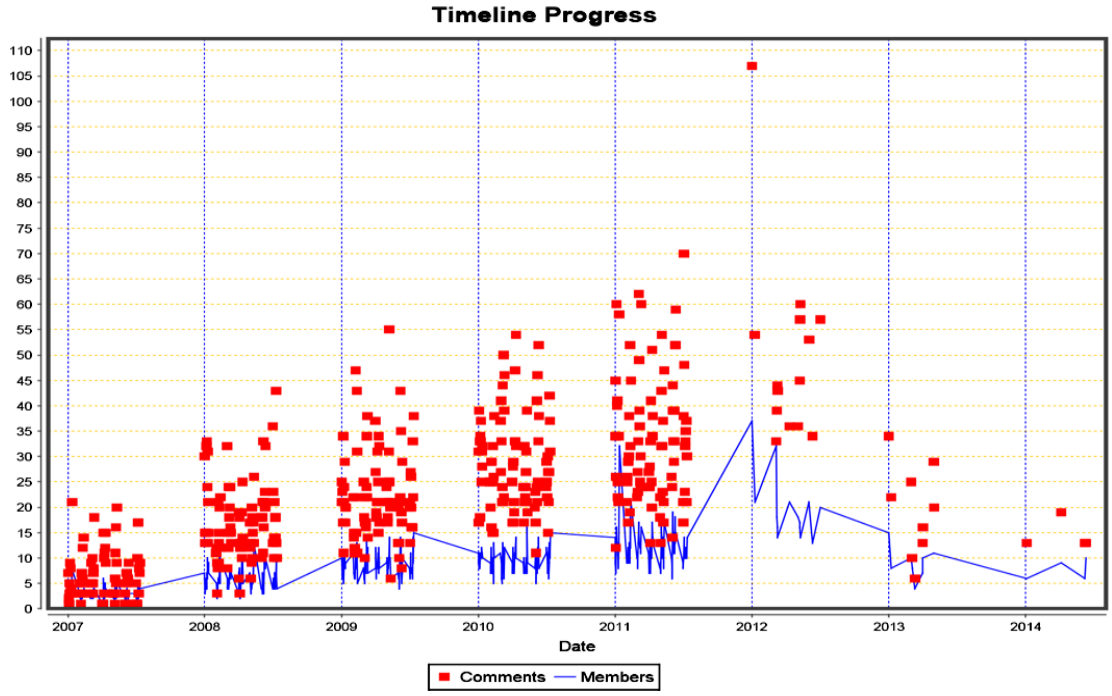
Fig. 3 depicts three different visualization scenarios. In all the examples, the x-axis quantifies time, while the y-axis represents various indices exposed by BlogCrawl. Fig. 3.a shows the evolution of the ratio of a post's number of comments and participants to the sentiment associated to the post (scale is [0-4], where 2 is neutral and 4 is very positive), in tight correlation with the main topics of the conversation. The second visualization (see Fig. 3.b) showcases a community in which a small group of members generates high inter-animation threads for 5 years followed by a sudden stop. The third evolution graph (see Fig.

3.c) highlights interaction patterns within a community, by following all members since their enrollment/first post up until their last contribution. Although the visualization is initially hard to follow, filtering members based on a minimum number of contributions enables the exploration of interaction patterns [11].

a)



b)



c)

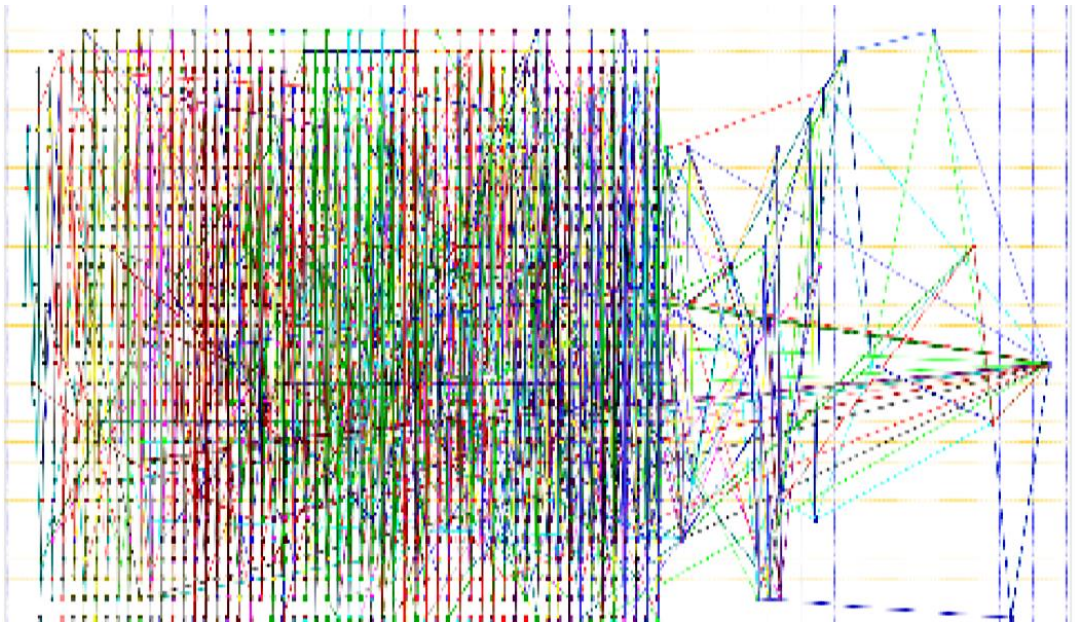


Fig. 3. Different visualization models

Integration with ReaderBench

ReaderBench [5, 12, 13] is an automated linguistic analysis framework based on advanced Natural Language Processing (NLP) techniques [14], that provides language support for Romanian [15], English [13, 16], French [8, 17], while Italian, Spanish and Dutch are currently under development. ReaderBench comprises methods for automated essay scoring [16], reading strategies identification, comprehension[12], discourse structure, CSCL, polyphony, and topic mining [13]. BlogCrawl's integration with ReaderBench [8] adds a linguistic dimension to the analysis by incorporating NLP techniques, assessment of participation and collaboration in the style of CSCL [18], and discourse structure in a single, comprehensive approach. Fig. 4 offers a sample visualization of the results obtained by using Cohesion Network Analysis (CNA), an in-depth assessment model of participation embedded into ReaderBench [13], on top of BlogCrawl data (members, posts, comments).

In sum, we must emphasize the profound *customizations* performed within BlogCrawl whose virtual DOM output representation, besides data pre-processing and different visualization options, greatly facilitate follow-up analyses performed within the ReaderBench framework.

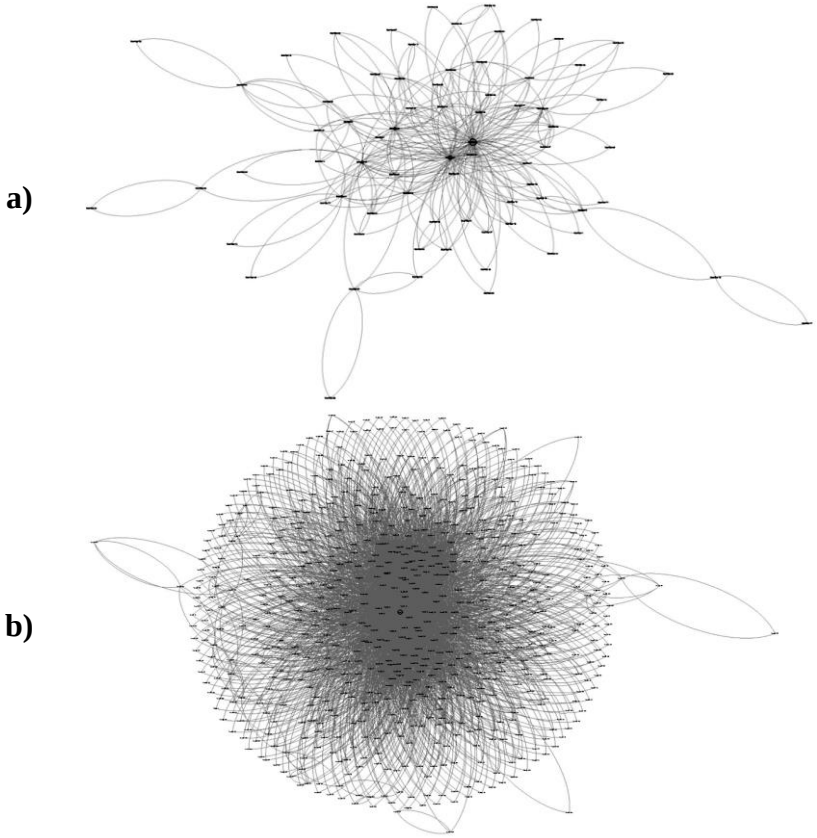


Fig. 4. Interaction graph samples corresponding to integrative (a) and non-integrative (b) communities

3. Case Studies and Results

BlogCrawl is an academic crawler embedding modules for crawling, parsing, data normalization, sentiment analysis, topic modeling, and timeline evolution analysis. While dedicated to building structured corpora for online collaborative environments like forums, chats, MOOCs, virtual communities of practice (vCOPs) or online knowledge communities (OKCs), it also provides integration with social platforms like Twitter, Facebook, or LinkedIn, and is generally compatible with platforms that expose one of the supported formats: XML, CSV, HTML, TXT or Excel.

Based on its internal representation of the conversational structure, the BlogCrawl framework implements metrics such as the number, length, and frequency of posts and comments, the degree of inter-animation, topic coverage, the sentiment associated with posts, comments, and topics, user contribution, and relationships between users. Also, the crawler makes it possible to study the evolution of these indicators over a specified amount of time. Additional analyses

come from the integration with ReaderBench. These characterize the activity of a member in the community in terms of: number of interventions and interactions with other members, length and quality of these interventions, the achievement of knowledge building at a personal or social level, degree of voice inter-animation, the relation to other members manifested as the position in the community graph (in relation to Social Network Analysis indices of closeness, betweenness, eccentricity).

Several studies [5, 6, 7, 11, 19] used BlogCrawl in different collaborative educational scenarios. Based on the metrics described above, a number of studies [11, 19, 20] classified OKCs into integrative and non-integrative communities. Integrative communities are characterized by fast and easy integration of new members, and encourage opinion sharing by all participants. Non-integrative communities are „moderated" by a small number of central users, who decide the acceptance or rejection of new members, and build knowledge that becomes characteristic of the entire community.

A quantitative analysis followed by a timeline analysis of one integrative and one non-integrative community [20] revealed significant differences in ratio of members per post and per comment, dynamism, or degree of interaction between members. Stavarache, Dascalu, Trausan-Matu & Nistor [19] selected 10 integrative and 10 non-integrative communities (based on human assessment), so as to meet the following criteria: at least 3 years lifetime, regular posting frequency, and a critical mass of members (over 50 members each). The purpose of the study was to identify the individuating traits (variables produced by using ReaderBench on top of BlogCrawl) of opinion leaders in OKCs, and to use the behavior of opinion leaders to predict the integrative/non-integrative character of a community. Opinion leaders are described as members with good reputation inside and outside the community, and representative voices for the community trends at any given time. The study departed from the presumption that opinion leaders are characterized mostly by intensive participation in the community, but discovered the number of contributions of a member to be a weaker predictor for the leader status than other indices like social knowledge building, closeness, eccentricity, or topics coverage.

In order to provide a fine-grained view, Stavarache, Dascalu, Trausan-Matu & Nistor [11] performed a side-by-side analysis of one integrative (politics) and one non-integrative (cooking) community over several 6 months intervals, with the purpose of identifying the main factors that determine the communities to expand or lose members. Using BlogCrawl and ReaderBench, sentiment polarity was extracted only in relation to the main topics addressed in the community, and only taking into account posts that generated at least one comment referring at least one main topic. The study outlined the similarities and differences between integrative and non-integrative communities. The first cover more topics and

consequently build knowledge faster. Sentiments also change faster in integrative communities, while there is a strong correlation in both communities between sentiment polarity and fluctuations in activity.

In addition, Nistor, Dascalu, Stavarache, Serafin & Trausan-Matu [6] ran a study with over 68 blogger communities, randomly selected from the Internet, to observe how a community's response to visitor inquiries varies with each of the following four factors: (a) the inquiry format (either on-topic or off-topic), (b) the topic of the blog, (c) collaborative dialog quality (assessed using ReaderBench on BlogCrawl's output), and (d) socio-cognitive structure. They found the response to be significantly influenced by the format of the inquiry (in cooking blogs), and by the collaborative dialog quality (in politics and economics blogs), while the community structure seemed to directly influence only the quality of the dialog, not the community response itself. The collaborative dialog quality was proposed as a predictor for a community's likelihood to be integrative and responsive.

Following a different path than the previous case studies, an analysis of the similarities and differences between Computer Supported Collaborative Learning papers and their corresponding slides [21] used BlogCrawl to clean (remove images, quotes, references) and normalize the content of the slides.

4. Conclusions and Future Work

Considering the large set of existing software tools for crawling, spidering, and sniffing, we introduce BlogCrawl as an integrated model of analysis for collaborative educational environments, that targets virtual communities of practice, forums, chats, and MOOCs. BlogCrawl offers an automated comprehensive model of crawling data from online educational and learning environments, normalizing it, and mapping the result onto the same standard structure, regardless of the original data source (see Fig. 2), a facility that is missing from other crawling mechanisms. The crawler further differentiates itself from other products through its compatibility with the ReaderBench framework, thus combining automated text complexity analysis, NLP techniques, and CSCL theories with processes of data extraction and validation from unsupervised environments.

BlogCrawl exposes how, when, and why the knowledge building process occurs outside the traditional educational setup of the tutor-student relationship. Furthermore, its timeline analysis (enhanced with topic detection and opinion mining capabilities) reflects the state of the learning communities at any given moment in time.

We foresee two main directions for further development: first, a quantitative expansion in terms of supported data sources and visualization scenarios; second, a qualitative refinement, by integrating new linguistic analyses

in ReaderBench, such as metrics of textual rhythmicity [22], rhetorical relation annotations, or methods for stimulating creativity [23, 24].

Acknowledgments

The work presented in this paper was partially funded by the EC H2020 project RAGE (Realising an Applied Gaming Eco-System) http://www.rageproject.eu/ Grant agreement No 644187.

R E F E R E N C E S

[1]. *E. Wenger*, Communities of practice. Learning, meaning, and identity (Learning in doing: Social, cognitive and computational perspectives), Cambridge University Press, Cambridge, UK, 1999.

[2]. *D. R. Garrison, T. Anderson and W. Archer*, "The first decade of the community of inquiry framework: A retrospective", in Internet and Higher Education, **vol. 13**, no. 1-2, 2010, pp. 5–9.

[3]. *Y. Wang*, "MOOC Leaner Motivation and Learning Pattern Discovery", in proceedings of the 7th Int. Conf. on Educational Data Mining, London, UK, pp. 452–454, 2014.

[4]. *G. Stahl*, Group cognition. Computer support for building collaborative knowledge, MIT Press, Cambridge, MA, 2006.

[5]. *M. Dascalu, L. L. Stavarache, S. Trausan-Matu, P. Dessus, M. Bianco and D. S. McNamara*, "ReaderBench: An Integrated Tool Supporting both Individual and Collaborative Learning", in proceedings of the 5th Int. Learning Analytics & Knowledge Conf. (LAK'15), Poughkeepsie, NY, ACM, pp. 436-437, 2015.

[6]. *N. Nistor, M. Dascalu, L. L. Stavarache, Y. Serafin and S. Trausan-Matu*, "Informal Learning in Online Knowledge Communities: Predicting Community Response to Visitor Inquiries", in proceedings of the 10th European Conf. on Technology Enhanced Learning, Toledo, Spain, Springer, pp. 447–452, 2015.

[7]. *N. Nistor, M. Dascalu and S. Trausan-Matu*, "Newcomer Integration in Online Knowledge Communities: Exploring the Role of Dialogic Textual Complexity", in proceedings of the 12th Int. Conf. on Learning Sciences (ICLS 2016), Singapore, International Society of the Learning Sciences (ISLS), pp. 914–917, 2016.

[8]. *M. Dascalu, P. Dessus, S. Trausan-Matu, M. Bianco and A. Nardy*, "ReaderBench, an environment for analyzing text complexity and reading strategies", in proceedings of the 16th Int. Conf. on Artificial Intelligence in Education (AIED 2013), Memphis, USA, Springer, pp. 379–388, 2013.

[9]. *C. D. Manning, H. Schütze*, Foundations of statistical Natural Language Processing, MIT Press, Cambridge, MA, 1999.

[10]. *D. Jurafsky, J. H. Martin*, An introduction to Natural Language Processing. Computational linguistics, and speech recognition, Pearson Prentice Hall, London, 2009.

[11]. *L. L. Stavarache, M. Dascalu, S. Trausan-Matu and N. Nistor*, "Topics Evolution in Online Knowledge Building Communities: A case study of cooking versus politics blogs", in proceedings of the 2nd Int. Workshop on Design and Spontaneity in Computer-Supported Collaborative Learning (DS-CSCL-2015), in conjunction with the 20th Int. Conf. on Control Systems and Computer Science (CSCS20), Bucharest, Romania, IEEE, pp. 765–772, 2015.

[12]. *M. Dascalu, P. Dessus, M. Bianco, S. Trausan-Matu and A. Nardy*, Mining texts, learner productions and strategies with ReaderBench, in Educational Data Mining: Applications and Trends, A. Peña-Ayala Ed. Springer, Cham, Switzerland, 345–377, 2014.

[13]. *M. Dascalu, S. Trausan-Matu, D. S. McNamara and P. Dessus*, "ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism", in International Journal of Computer-Supported Collaborative Learning, **vol. 10**, no. 4, 2015, pp. 395–423.

[14]. *C. D. Manning, M. Surdeanu, J. Bauer, J. Finkel, S. J. Bethard and D. McClosky*, "The Stanford CoreNLP Natural Language Processing Toolkit", in proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations, Baltimore, MA, ACL, pp. 55–60, 2014.

[15]. *D. Gifu, M. Dascalu, S. Trausan-Matu and L. K. Allen*, "Time Evolution of Writing Styles in Romanian Language", in proceedings of the 28th Int. Conf. on Tools with Artificial Intelligence (ICTAI 2016), San Jose, CA, IEEE, pp. 1048–1054, 2016.

[16]. *L. K. Allen, M. Dascalu, D. S. McNamara, S. Crossley and S. Trausan-Matu*, "Modeling Individual Differences among Writers Using ReaderBench", in proceedings of the 8th Int. Conf. on Education and New Learning Technologies (EduLearn16), Barcelona, Spain, IATED, pp. 5269–5279, 2016.

[17]. *M. Dascalu, L. L. Stavarache, P. Dessus, S. Trausan-Matu, D. S. McNamara and M. Bianco*, "Predicting Comprehension from Students' Summaries", in proceedings of the 17th Int. Conf. on Artificial Intelligence in Education (AIED 2015), Madrid, Spain, Springer, pp. 95–104, 2015.

[18]. *T. Koschmann*, "Toward a dialogic theory of learning: Bakhtin's contribution to understanding learning in settings of collaboration", in proceedings of the Int. Conf. on Computer Support for Collaborative Learning (CSCL'99), Palo Alto, ISLS, pp. 308–313, 1999.

[19]. *L. L. Stavarache, M. Dascalu, S. Trausan-Matu and N. Nistor*, "Predicting the Integration of Newcomers in OKBCs based on Existing Members' Involvement", in proceedings of the 6th Int. Conf. on Information, Intelligence, Systems and Applications (IISA 2015), Corfu, Greece, IEEE, pp. 1–5, 2015.

[20]. *L. L. Stavarache, M. Dascalu, S. Trausan-Matu and N. Nistor*, "How Does Time Shape a Virtual Community of Practice?", in proceedings of the 4th Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 11th Int. Conf. on eLearning and Software for Education (eLSE 2015), Bucharest, Romania, Carol I NDU Publishing House, pp. 380–386, 2015.

[21]. *L. L. Stavarache, M. Dascalu, S. Trausan-Matu and P. Dessus*, "Papers vs. Slides: Do they have similar textual traits?", in proceedings of the 3rd Int. Workshop on Semantic and Collaborative Technologies for the Web, in conjunction with the 10th Int. Conf. on eLearning and software for Education, Bucharest, Romania, Editura Universitatii Nationale de Aparare "Carol I", pp. 187–192, 2014.

[22]. *M. Balint, M. Dascalu and S. Trausan-Matu*, "Classifying Written Texts through Rhythmic Features", in proceedings of the 15th Int. Conf. on Artificial Intelligence: Methodology, Systems, and Applications (AIMSA 2016), Varna, Bulgaria, Springer, pp. 121–129, 2016.

[23]. *D. Stamati, M. Dascalu and S. Trausan-Matu*, "Creativity stimulation in chat conversations through morphological analysis", in Scientific Bulletin, University Politehnica of Bucharest, Series C, **vol. 77**, no. 4, 2015, pp. 17–30.

[24]. *A. Oprisan, S. Trausan-Matu*, "Creativity Stimulation Tool", in Annals of the Academy of Romanian Scientists, Series on Science and Technology of Information, **vol. 6**, no. 1, 2013, pp. 63–83.