Building a Semantic Recommendation Engine for News Feeds based on Emerging Topics from Tweets

Mihai Tabara, Mihai Dascalu, Stefan Trausan-Matu
University Politehnica of Bucharest, Computer Science Department
Bucharest, Romania
tabara.mihai@gmail.com, mihai.dascalu@cs.pub.ro, stefan.trausan@cs.pub.ro

***Abstract*—The rise of social networks powered by the emergence of Web 2.0 unleashed a massive amount of generated user content. Concurrently with technology enhancements that facilitated its widespread, Web 2.0 became the engine which hastened the appearance of worldwide mass communication techniques. Alongside its advent, textual analysis changed as new user-centered content failed to comply with traditional grammar ruling. In this paper, we approach the problem of topic extraction from Twitter in the context of designing a recommendation engine to best matching user profiles to news feed articles. We propose a strategy to extract the concepts by means of Natural Language Processing and use of the semantic cohesion measurements to leverage the matching process. In order to prove the adequacy of our method, we have conducted a medium-scale evaluation. Our results demonstrate the particularities of the Twitter textual corpora, as well as how it can be used to infer geo-locations for its users.**

Keywords—topic extraction, news recommendation, tweets, prediction of news feeds.

I. Introduction

Automated textual analysis has been of particular interest ever since the appearance of the computers in the modern technology era. While traditional texts consisting of novels, poems or stories have served well in text analysis, a new form of much more dynamic content emerged within the Web 2.0 [1]. The grammatical rules and lexical constructions are not always respected, sentences started to change their traditional structure and people migrated towards comfortably minimizing the typed information in their attempt to maximize the meaning of the transmitted message. Under these conditions, the text boundaries have been broken and various types of data and meta-data emerged, enriching the context of semantic representation. Text analysis became more challenging, but also more feasible, as computational capabilities have significantly grown in the past decades.

Twitter is one of the most widely and frequently used social platforms that emerged within the social web due to its capabilities. Shortly put, Twitter provides the hammer and nails for hosting a micro-blogging platform. In colloquial language, users become "*twitterers*" and are automatically assigned a "*wall*" on which they can post various thoughts, can express feelings or share interesting artifacts. In contrast to other platforms, Twitter limits the number of characters written to 140. Though initially considered a severe limitation, it was quickly acknowledged as a successful tip to enforce the quintessence of information across its network [2]. The limitation itself has drawn attention to researchers as it encompasses, at its core, a huge potential for analyzing people's behavior, as well as their capabilities of expressing the essential information.

Topic mining has emerged as a viable automated artificial intelligence technique which is used to detect the most relevant concepts from a given text input [3]. Unfortunately, the task at hand increases in difficulty for loosely-coupled structures such as word sequences (i.e., text segments with few tokens created and employed by human individuals). In this paper, we use various Natural Language Processing (NLP) techniques, including semantic distances in lexicalized ontologies, Latent Semantic Analysis (LSA) vector models and Latent Dirichlet Allocation (LDA) topic distributions, all integrated within our *ReaderBench* framework [4-7] in order to extract the key concepts from a Tweet corpus. Furthermore, we attempt to match valid news feeds extracted via crawling to two major news broadcasting systems: CNN and BBC.

This section describes the general problem of topic extraction of content from Twitter, as well as the problematic matching between user profiles and news feeds. Topic extraction is a vast study domain and can be applied to a large variety of documents, including the Twitter corpora. The specificity of this input is given by the unique attributes of a tweet: its limited length and its colloquial content. Given a set of world-wide distributed Twitter users, each with at least 2000-3000 messages posted on the Twitter social network, commonly referred to as *tweets,* the proposed solution aims to extract the main concepts and topics. With the resulting topics, a matching between the user concept-profile and news feed was created. In addition, the problem of concept extraction can be further divided into multiple sub-tasks:

- the textual corpus needs to be consistent and broad in terms of the coverage of underlying concepts and phrases;
- meaningless words such as the function words, as well as other named entities or non-lexical elements (links, media images, etc.) need to be filtering out;
- a specific matching algorithm needs to be designed in order to facilitate concept extraction.

The first task is relatively simple as tweets are individual comprehensive pieces of text that a user shares on the social network. Independently, each tweet usually denotes a different feeling or state and can vary in its form, ranging from personal

quotes or a shared news, to social events sharing. While each tweet taken individually provides very little data for processing collections of concatenated tweets create a powerful and meaningful text corpus. The second task refers to the pre-processing of the crawled data which can be performed using a Natural Language Processing pipeline [8]. Lastly, the subset of concepts needs to be extracted. The matching of user Twitter profiles with real news feed is a follow-up task as main concepts and topics excerpted from the document corpora are compared against a similar subset obtained from the news feed. By means of semantic similarity and cohesion, this paper focuses on presenting such results relying on CNN and BBC broadcast news feeds.

Section II describes recent approaches for topic mining in tight correlation with Twitter features, as well as its limitations. Mining techniques and main architecture are presented in section III, while section IV covers the results, followed by conclusions.

II. State of Art

Understanding the particularities of human communication, regardless of the language itself, has been conducted in many forms for centuries. Nowadays, with the burst of modern communication, this task has grown in importance. However, some of the historically assumed methods are still valid today l. Such an example is the use of a thesaurus (2001-2016) [9], which is a special kind of a dictionary that groups words according to their similarity of meaning. Potential concepts can be inferred from a group of synonyms or words sharing the same lexical and semantic field [10]. A more recent approach is Princeton's WordNet which consists of an English lexical storage base [11]. The most challenging piece in topic mining is the mapping of words to concepts. As human are able to solve ambiguity by relating to context, this task tends to be difficult for the machines as inferring context is not trivial. However, recent advanced techniques such as reckoning semantic similarities can be used to address this problem at large scale.

Topic mining usually implies large text corpora. Starting from a large collection of documents, various techniques can be used in order to achieve disambiguation. A text analysis can be conducted to extract the parts of speech, followed by a frequency-based mapping between words and concepts. For large quantity of data, this approach infers from context and is able to validate potential disambiguation. Because of its inner features, a Twitter-based text corpus encounters other challenges as it allows its users to share short text messages, not longer than 140 characters. It can serve many purposes that vary from breaking news to sharing URLs or social tool usage, to inane comments tweets or, more commonly, sharing personal updates and spontaneous opinions. Due to its nature of micro-blogging service, the large amount of text may presumably contain useful information in personalizing user interests. This section describes some of the recent approaches for topic mining, as well as its coping purpose of addressing Twitter features and limitations. Furthermore, we present results from different papers that have attempted to best match Twitter profiles with news feeds. Moreover, we look for results on unsupervised learning classifications based on clusters within the Twitter data.

A. Topic Mining

Given the limitation of a tweet text, one may assume that ideas are expressed within a higher grade of granularity as, usually, the reduced number of characters allow for uttering a statement are related to a single category. Because of this, the foundation for all text mining strategies lies within the documents' representation. Elkan [12] states that the most commonly used document representation is a two-dimensional term-document matrix (each row describing a document, each column corresponding to a word) used for subsequent low-rank approximations. Given a large text corpus, one must determine the vocabulary defined within the set. Subsequently, an iteration is performed across all documents to select those words that are present in at least two documents. Out of the words contained in the vocabulary, stop-words need to be disregarding as they do not provide any content: pronouns (e.g., "you", "he"), prepositions (e.g. "on", "after"), articles or discourse connectives. While the document representation provides a way to technically structure the information, a model ensures a more comprehensive abstraction.

Elkan describes the difference between a model and a representation as follows: "a representation is a way of encoding an entity as a data structure whereas a model is an abstraction of a set of entities, for example, a probability distribution" [12]. Thus, modeling topics across the set of documents requires a multinomial probability distribution. Elkan then proposes a generative process followed by Latent Dirichlet Allocation (LDA) [13, 14], an unsupervised learning approach based on a generative probabilistic process. The training set of documents will eventually cope with the values chosen for the probabilistic model to provide higher probability.

While many researchers rely on traditional topic mining techniques to understand Twitter content, Hong et al. [15] chose to set the *140ch*-length limitation as a deal-breaker and propose an unconventional approach instead, based on standard topic models in the micro-blogging environment. Hong et al. train the standard topic model throughout several systems by assessing their behavior within already certified prospects. After aggregating all the user messages, they divide the content in training and testing profiles, followed by the inference of topic mixtures. According to their study, a higher quality learning model can be achieved by training a topic model on aggregated short messages, as the efficacy of trained topic models is heavily affected by the length of the documents.

B. User/news matching

Porrata et al. [16] present a topic discovery system aimed to reveal the implicit knowledge from news streams. They propose a topic hierarchical model in which each topic contains a subset of documents relating to it, as well as a brief abstract. Their follow-up consists of a new incremental hierarchical clustering algorithm, which reunites both partitioned and agglomerating approaches, as well as a new summarization method to build the topic summaries.

Cataldi et al. [17] have approached the problem from a different angle. Rather than focusing on the content itself, they notice the particularities of the Twitter network in which users of various ages and social conditions are widely spread across the world. The emphasis falls on Twitter's short response time, hence its inborn news-portal condition. Focusing on this model, the geometry changes as novelty becomes the topic detection technique, with all its subdivisions. The authors look at the set of terms within the Twitter content to formalize a novel aging theory in order to detect the rising terms or trends. Cataldi et al. [17] consider a term as being *emergent* if "it frequently occurs in the specified time interval and it was relatively rare in the past". To enhance their approach even more, along with frequency, their analysis also encompasses the source of the content and the authority of the users by means of Page Rank [18]. To validate their approach, a topic graph is used to connect the emerging concepts with semantically related subtopics.

Zhao et al. [19] present a solution that highly resembles the one presented in this study. The authors question Twitter's news-portal nature as it is not yet validated, whether it is a generative news feed ecosystem or just an environment that enhances traditional news media propagation. Similar to our approach, Zhao et al. compare the content from Twitter to a traditional and worldwide known news stream - the New York Times. Using unsupervised topic modeling and various text mining techniques such as LDA, Zhao et al compare the discovered topics from Twitter content with categories and labels from New York Times. Furthermore, they measure the extent of the opinionated tweets and retweets.

Dredze et al. [20] take the analysis even further into mining the Twitter corpus for public health measures by attempting to correlate Twitter content messages with flu rates in the United States. By relying on an Ailment Topic Aspect Model [21], Dredze et al. extract a high number of pathologies including maladies, obesity, and insomnia. By mining public health information both over time (e.g., syndromic surveillance for influenza) and by geographic region (e.g., behavioral risk factors), they are able to track down common ailments such as seasonal allergies and prove known patterns (e.g., a higher frequency during Spring). The primary allergen in the spring is tree pollen, indicated by the presence of "pollen" and "trees" in April. By applying various heuristics such as localizing illnesses by geographic region or analyzing symptoms of medication usage, Dredze et al. prove that Twitter data has a large applicability into health research.

III. System's Architecture

People have the tendency to express mood swings and mixed feelings, thus being rather emotional or negative on Twitter. All these feelings and states, along with non-emotional content information such as news, create an appealing personalized data source accessible via the Twitter API. A dedicated crawler was used to gather and store the required data. As part of the data was well nested into the JSON response object, a NoSQL database (i.e., MongoDB) was particularly useful for simplifying tweet storage; no tweet parsing or relationship mapper were required at this stage. Unlike strongly typed relational databases, the NoSQL document database has information from the data itself and, most importantly, allows every data instance to be different from any other. In order to ensure a comprehensive corpus, a strong collection of users alongside the CNN Breaking News official twitter account were used, resulting in 10K seeded tweets, 120K retweets, 10K users and ~2M individual tweets.

Starting from the previous text collection, our model makes use of the *ReaderBench* framework, a multi-purpose, multilingual tool that creates a powerful environment to support reading and writing activities for both learners and teachers alike, each learning task having its specificities and particularities. According to our previously conducted experiments [6, 7, 22], *ReaderBench* facilitates the automated assessment of three main textual specificities, all thoroughly validated: cohesion-based assessment, reading strategies identification and textual complexity evaluation.

Our proposed approach uses *ReaderBench* as a text mining engine. Through *ReaderBench*, the input tweets suffered filtering techniques that consisted of tokenization, reducing inflected words to their word lemma, stop words removal, tagging across parts of speech, parsing, named entity recognition, and co-reference resolution. In contrast to the previously described methods, *ReaderBench* makes use of three complementary semantic models in order to evaluate text cohesion [6]: semantic distances in WordNet [23] Latent Semantic Analysis (LSA) vector models [24] and Latent Dirichlet Allocation (LDA) topic distributions. Thus, ReaderBench makes uses of Cohesion Network Analysis to provide an in-depth discourse structure centered on cohesion [25]. All topics were extracted from the already processed tweet corpora and sorted out by relevance. *ReaderBench* uses its multi-layered cohesion graph to represent the general structure of all tweets corresponding to a specific user. The tweets represent the central constituents of the artificially created document and they are mapped to the blocks within our generalizable model, despite their obvious length limitation.

Upon extracting the topics for each individual, an additional corpus of RSS news feed scraped from CNN was fed to *ReaderBench* in order to compute semantic similarity between tweets and news articles. Each paragraph from the news articles maps to a generic block from our semantic engine. A similar approach to the analysis of tweets corpora is adopted, hence consisting of pre-processing techniques, followed by parsing and content words filtering. Once the news articles share the same semantic representation with the tweets (i.e., word vectors within the semantic models integrated into *ReaderBench*), we determine the best matching news category for each Twitter user by relying on a greedy approach [26] that recommends news categories based on the highest score of semantic similarity. Although being representative for certain sub-domains, common words across multiple news categories, were disregarded.

Figure 1 depicts the system's architecture and highlights all underlying information flows and processes. A clear demarcation has been made between the *ReaderBench* framework which was previously developed and which was

integrated for this study, in contrast to the new modules that were specifically implemented.

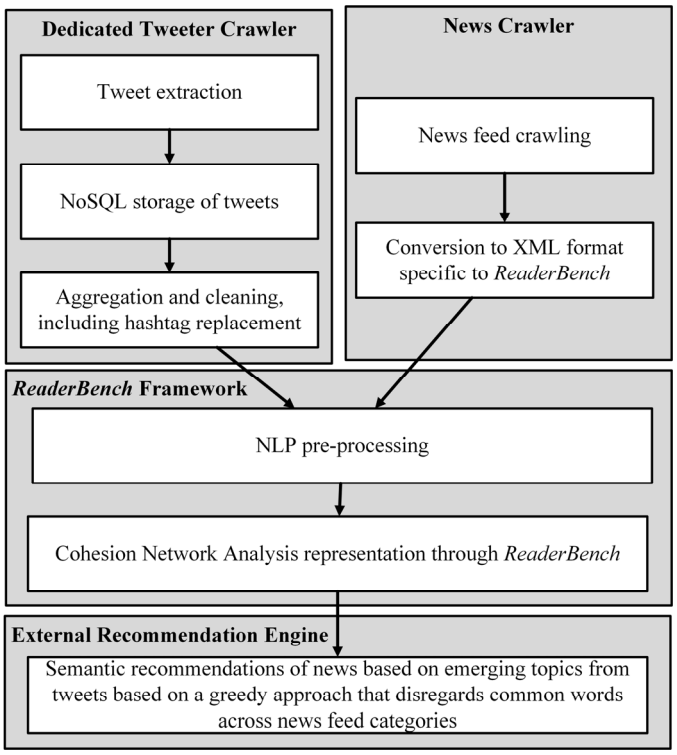

Fig. 1. System's architecture and the information flow.

IV. Results

Our validation experiment consisted of an in-depth analysis of a subset of 21 Twitter profiles (see Table 1 for descriptive), each with a full Twitter timeline excerpted from our 2 million tweets corpus. The selection criterion considers the coverage of all major geographical regions (11 from both Americas, 2 from Africa, 2 from Australia, 5 from Europe, 1 from Asia), gender equitability, a wide range of friends/followers/tweets, as well as a large age spectrum that varies from 16 to 55.

TABLE I. General statistics about the considered individuals.

	Min	Max	Avg	Stdev
Friends	41	29415	2,282.67	6,310.11
Followers	20	29476	2,858.90	6,698.56
Favourites	9	33397	5,621.90	8,843.77
Posts	26	53761	11,563.71	14,061.90

The information from the news feeds was organized across several categories that aggregated all corresponding articles. The semantic similarity was computed as the average value of WordNet, LSA, and LDA semantic similarities [6] between all the posts of each user and all news articles within a given category. In order to refine our results and create a better differentiation among categories, we have opted to remove the common words pertaining to all categories; thus, the values dropped by an average 10%. However, table 2 depicts high values of average semantic similarity across all categories, thus indicating a wide spread of interests, as concepts from all categories are conveyed. These results show that, even after eliminating the most common concepts across all news feed categories, the vocabulary used in Twitter is still colloquial. The "*America*" topic covering events that occurred within specified region was the most predominant in user preferences, followed by "*entertainment*", "*sport*" and "*tech*" categories.

TABLE II. Average semantic similarities between news categories and user's tweets.

Category	Average	Standard deviation
America	.72	.05
Entertainment	.69	.04
Sport	.66	.05
Tech	.66	.05
Asia	.66	.05
Arts	.65	.04
Travel	.63	.05
Europe	.62	.05
Africa	.62	.05
Middle East	.57	.05
Money	.54	.04
Music	.52	.03
Nature	.50	.04

In addition, Figure 2 depicts a dendrogram generated after considering an agglomerative clustering algorithm that uses the Pearson correlations between the semantic similarities of each user with all existing categories. An interesting element is that the most semantically related user profiles originated from the same continent (America, having the following user IDs: 1, 5, 9, 11, 12, 15, 19).

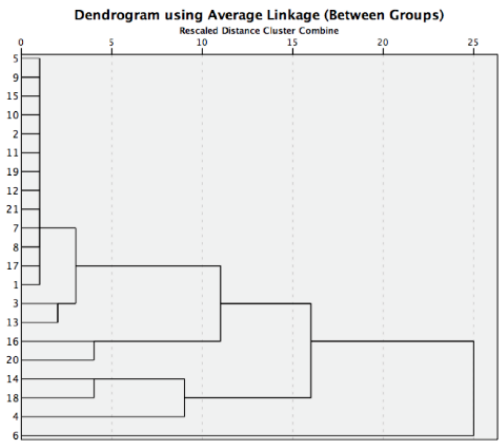

Fig. 2. Dendrogram highlighting user similarities in terms of categories of interest.

V. Conclusions

The more social networks evolve, the more human interaction behaviors change due to technology enhancements; thus, new research methods are needed in order to understand this microform of communication. This paper presented various solutions encompassed in the topic extraction problem, as well as techniques used to best match news feed with Twitter user profiles. Our main focus was to rely on

various semantic models (WordNet, LSA, and LDA) integrated into our *ReaderBench* framework in order to extract the main concepts from a subset of Twitter text corpora. The resulting concepts are compared with a collection of news feed articles in terms of semantic similarity. Though results indicated at first glance similar similarity values across all categories, further evaluations shed light on the social nature of Twitter, where the used vocabulary is colloquial, therefore intrinsically implying high correlations across categories.

This study extends the functionalities of our framework and provides a semantic approach suitable for recommending relevant articles starting from users' tweet messages. We introduce an extensible approach that takes into consideration the limitations of tweets and applies subsequent optimizations in order to provide specific recommendations. In addition to our previous studies centered on retrieving semantically related articles based on user queries [27, 28], we shift the focus towards a different outcome by integrating new data sources, by creating user profiles based on topics, and by providing a comprehensive processing and matching approach that disregards common words.

Acknowledgment

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

References

[1] Ed H. Chi, "The Social Web: Research and Opportunities," Computer 41 (9), pp. 88–91, 2008.

[2] B. J. Jansen, M. Zhang, K. Sobel, and A. Chowdury, "Twitter power: Tweets as electronic word of mouth," Journal of the American society for information science and technology, 60(11), pp. 2169-2188, 2009.

[3] B. Pang, and L. Lee, "Opinion mining and sentiment analysis," Foundations and Trends in Information Retrieval, 2(1-2), pp. 1–135, 2009.

[4] M. Dascalu, L. L. Stavarache, P. Dessus, S. Trausan-Matu, D. S. McNamara, and M. Bianco, "ReaderBench: An Integrated Cohesion-Centered Framework," in 10th European Conf. on Technology Enhanced Learning, Toledo, Spain, 2015, pp. 505–508.

[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 5th Int. Learning Analytics & Knowledge Conf. (LAK'15), Poughkeepsie, NY, 2015, pp. 436-437.

[6] M. Dascalu, Analyzing discourse and text complexity for learning and collaborating, Studies in Computational Intelligence vol. 534. Cham, Switzerland: Springer, 2014.

[7] 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., ed Cham, Switzerland, Springer, 2014, pp. 345–377.

[8] D. Jurafsky, and J. H. Martin, "An introduction to Natural Language Processing. Computational linguistics, and speech recognition," 2nd ed. London, Pearson Prentice Hall, 2009.

[9] D. Harper, "Online Etymology Dictionary", 2001-2015, http://www.etymonline.com/index.php?term=thesaurus

[10] R. Peter, "Thesaurus of English Language Words and Phrases," Available online at: http://www.thesaurus.com/roget/ (Retrieved August 10, 2016).

[11] G. A. Miller, "WordNet: A lexical database for English," Communications of the ACM, vol. 38, pp. 39–41, 1995.

[12] C. Elkan, "Text mining and topic models", February 12, 2014, pp. 1-6.

[13] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet Allocation," Journal of Machine Learning Research, vol. 3, pp. 993–1022, 2003.

[14] D. M. Blei, and J. Lafferty, "Topic Models," in Text Mining: Classification, Clustering, and Applications, A. Srivastava and M. Sahami, Eds., ed London, UK: Chapman & Hall/CRC, 2009, pp. 71–93.

[15] L. Hong, D. Davison, "Empirical study of topic modeling in Twitter", Proceedings of the First Workshop on Social Media Analytics, 2010, 80-88.

[16] A. P. Porrataa, R. Berlanga-Llavorib and J. Ruiz-Shulcloper, "Topic discovery based on text mining techniques", 2006, pp. 1-4.

[17] M. Cataldi, L. Di Caro, and C. Schifanella, "Emerging topic detection on Twitter based on temporal and social terms evaluation", in proceedings of the 10th International Workshop on Multimedia Data Mining (MDMKDD'10), 2010.

[18] L. Page, "Method for node ranking in a linked database," USA Patent 6,285,999, 2001.

[19] W.X. Zhao, J. Jiang, J. Weng, J. He, E. Lim, H Yan, and X Li, "Comparing Twitter and Traditional Media Using Topic Models", Singapore Management University, Singapore, 2011.

[20] M. J. Paul and M. Dredze, "You Are What You Tweet: Analyzing Twitter for Public Health", Human Language Technology Center of Excellence, Center for Language and Speech Processing, Johns Hopkins University, 2011, pp. 1-4.

[21] M. Paul, and M. Dredze, "A model for mining public health topics from twitter," Technical report, Johns Hopkins University, 2011.

[22] S. Trausan-Matu, M. Dascalu and P. Dessus, "Textual Complexity and Discourse Structure in Computer-Supported Collaborative Learning", In: 11th Int. Conf. on Intelligent Tutoring Systems, 2012, pp. 352–357.

[23] A. Budanitsky, and G. Hirst, "Evaluating WordNet-based Measures of Lexical Semantic Relatedness," Computational Linguistics, vol. 32, pp. 13–47, 2006.

[24] T. K. Landauer, and S. T. Dumais, "A solution to Plato's problem: the Latent Semantic Analysis theory of acquisition, induction and representation of knowledge," Psychological Review, vol. 104, pp. 211–240, 1997.

[25] M. Dascalu, S. Trausan-Matu, D. S. McNamara, and P. Dessus, "ReaderBench – Automated Evaluation of Collaboration based on Cohesion and Dialogism," International Journal of Computer-Supported Collaborative Learning, vol. 10, pp. 395–423, 2015.

[26] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, Eds., "Introduction to Algorithms". Cambridge, MA: MIT Press, 2009

[27] I. C. Paraschiv, M. Dascalu, P. Dessus, S. Trausan-Matu, and D. S. McNamara, "A Paper Recommendation System with ReaderBench: The Graphical Visualization of Semantically Related Papers and Concepts," in *State-of-the-Art and Future Directions of Smart Learning*. vol. Lecture Notes in Educational Technology, Y. Li, M. Chang, M. Kravcik, E. Popescu, R. Huang, Kinshuk, *et al.*, Eds., ed Berlin, Germany: Springer-Verlag Singapur, 2015, pp. 443–449.

[28] I. C. Paraschiv, M. Dascalu, S. Trausan-Matu, and P. Dessus, "Analyzing the Semantic Relatedness of Paper Abstracts - An Application to the Educational Research Field," in *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, 2015, pp. 759–764.