INTED 2017

11th International Technology, Education and Development Conference

6-8 March, 2017
Valencia (Spain)

CONFERENCE PROCEEDINGS

Sharing the Passion for Learning

Published by
IATED Academy
iated.org

INTED2017 Proceedings
11th International Technology, Education and Development Conference
March 6th-8th, 2017 — Valencia, Spain

Edited by
L. Gómez Chova, A. López Martínez, I. Candel Torres
IATED Academy

ISBN: 978-84-617-8491-2
ISSN: 2340-1079
Depósito Legal: V-369-2017

Book cover designed by
J.L. Bernat

ANALYSIS OF USER PROFILES IN SOCIAL NETWORKS TO SEARCH FOR PROMISING ENTRANTS

A. Feshchenko, V. Goiko, G. Mozhaeva, K. Shilyaev, A. Stepanenko

National Research Tomsk State University (RUSSIAN FEDERATION)

Abstract

Educational globalization makes leading universities search for new ways of recruitment aimed at gifted smart youth not only from the same but from other countries as well. Since the university resources are limited from the point of view of coverage and attraction of the entrants, there is a need for a focused informational influence on a precise audience with specific features. This audience consists of high school students who show some interest in certain academic subjects and have soft skills for a successful study and an academic career. At the same time, the “natural habitat” of the modern schoolchildren is social networks. These social networks are the source of open data basing on which one can define potential entrants with a set of the required features according to the university entrant model. These data analysis and interpretation let a university to find promising entrants in any region or even a country and establish a direct communication with them via the social networks without any mediators.

The current paper covers the experience in collecting and analyzing the data about the users of social networks (the Russian social network VKontakte is taken as an example) to define the potential entrants. The authors provide a solution to the tasks related to building an entrant model, exporting data from the social networks using API, processing natural language, defining the entrants’ soft skills and educational interest via the analysis of the data taken from their profile, their walls, their friendship connections.

Keywords: Social Media, data analysis, big data, natural language processing, attract entrants.

1 INTRODUCTION

Being one of the universities participating in “5-100” Russian Academic Excellence Project, National Research Tomsk State University concerns academic recruitment in Siberian region as one of the major objectives. Annually the general population of potential entrants amounts to more than 200 thousand people, spread over the territory of more than 5 million square kilometers. About 150 universities in the region compete for that audience. To solve the task of defining the most promising entrants and attracting them to TSU, the research team studies the opportunities of user data analysis derived from the social network «Вконтакте» (“VKontakte”), that is the largest and the most popular social network among Russian young people.

The purpose of the study is searching for the methods of precise detection of high school students with high educational achievements, developed soft skills and stable interest to a particular science among the social network users.

Our hypothesis is that a social network user leaves a digital footprint in the virtual reality. Searching and analyzing this footprint, we may reveal the user’s academic needs and predict his/ her ability to enter a university for a specific program. Testing of the hypotheses requires solving the following research objectives: building a TSU entrant target model, searching for the network users who are considered to be potential entrants, downloading their data from the open network sources and its processing and comparing the analysis results with the entrant target model to define the users of the most relevant models. The current paper is devoted to the results of solving the only one objective, that is downloading user data from the network and its linguistic analysis aimed at defining the academic interests.

As far as we are concerned, the interest of the high school students in one or another subject field is connected to the probability of entering a particular university faculty. A user’s interests in social networks are presented via the texts published on his/ her profile wall. Analyzing these texts, to our mind, we are able to define the research areas this user is inclined to. Then we can divide possible entrants according to their interest into three groups (Humanities, Life Science, Physics- Mathematics) and differentiate users in every group on the extent of their interest.

To test this idea we have selected the “Vkontakte” profiles of TSU entrants in 2016 who are currently studying at diverse university faculties. Therefore, we managed to compare the texts that users published before entering TSU with their educational interests and objectives expressed via entering a particular faculty.

2 METHODOLOGY

The research is based on content analysis, quantitative analysis, qualitative analysis, text mining, and automated text processing. The basic tool for mining data from social networks is Application programming interface (API). It enables getting public data including the fields in a user profile (name, surname, city, country, gender, education, interests, favorite books and so on), content of a user personal page (wall), and a list of friends and a user activity (likes, comments, reposts) in different communities. Searching for the freshmen’s “Vkontakte” accounts via API, we considered the lists from the university CRM (name, surname and faculty) as primary data. As a result, we managed to find accounts of 73% students that were the sources for the text data taken from the wall (posts). These posts have been published from September, 1 2014 to September, 1 2016, that is the period of time when our today university students were entrants.

The main data source for the analysis of entrants’ interests in the sphere of education was the wall – the public personal space in the user’s social network page. Various linguistic studies have focused on the wall as the expression of one’s Internet identity [1-4]. These works, as well as the majority of works that focus on personal linguistic features on the Internet, cannot avoid the question of how and to what degree the real-life identity and the virtual identity match each other. This problem is of special importance to our study, since our final goal is to locate the real prospective student. The works of M. Back, S. Gosling, A. Voiskunsky [5-7] show that the virtual (or alternative) identity of most people correlates closely with their self-presentation in real life. The validity of our study is further increased by automatically eliminating empty and mostly incomplete profiles of Vkontakte users.

The material for the linguistic component of our study was sourced from posts of different genres, the latter understood as forms of representation of textual information. Scholars note that the wall can be characterized by the presence of a number of genres, such as compliments, advertising, news or entertaining messages [8, 9, 4]. As a whole, the wall presents the features of virtual communication pointed out by numerous scholars [11,12]. Its hypertextual and interactive character is manifested in the high number of comments and reposts, which constitute the bulk of the content. The intensive usage of multimedia, often in the reposts themselves, and the growing presence of non-verbal elements (such as images, or the post consisting solely of an image) are also typical. The fact that the content of the wall is usually fragmented, the bits of text are short and repetitive, and the percentage of visual content is high, limits the applicability of grammatical, syntactic, pragmatic or cognitive text analysis, otherwise effective.

As regards the information content of the wall, the ratio of the author’s original content is relatively low: 4,7% of the total text on average (ranging from 1,8% to 9,8%). This relatively low percentage allows us to assume that the linguistic differences between the content of the user’s wall and the content of the community posts that are subsequently reposted are negligible. The resulting conclusion is that the dominant genre on the wall, which provides most of its textual content, is the community post.

We view the community post as the dominating genre in the user’s wall (the post as a genre was described by E. Goroshko [10]). The majority of the posts that served as material for automated content analysis are short messages of educational or entertaining nature. They are usually accompanied by a photograph or other image (drawing, infographics). The target audience is members of the community, who have already shown their interest in its topics by joining the group. Consequently, finding quantitative correspondence between the total texts of a user’s wall a community wall should reveal the user’s interests.

In order to find the correlation, we make use of content analysis. We follow D. Riffe et al. in their definition of its essence and aims: [Content analysis is defined as] «systematic and replicable examination of symbols of communication, which have been assigned numeric values according to valid measurement rules and the analysis of relationships involving those values using statistical methods, to describe the communication, draw inferences about its meaning, or infer from the communication to its context, both of production and consumption» [14]. By counting relevant textual units (in our case, these are separate subject-related words) we plan to discover the interest of a particular user – a potential student – towards a particular area of knowledge.

3 RESULTS

The content analysis was conducted in two stages: the analysis of textual content of thematic communities and the analysis of a potential user's wall. The array of individual messages on the user's wall or the community wall served as sample. The word was chosen as a unit of analysis. Quantitative analysis does not allow to make inferences ("read between the lines"), so another assumption was necessary for the study: the user "means what he says", i.e., the priority of literal meaning was assumed. While we did not create concordances and so did not take into account the context of the keywords, we believe that the strengths of the content analysis method compensate for this. It is capable of determining the thematic vector of the text [13] and it allows to analyze the natural conduct of the target audience and conduct longitudinal studies and post-tests with archived data [14]. The main reason for applying CA in our study is our aim to increase the scope of the entrants, whose digital footprint is nearly impossible to follow manually.

Research hypothesis was the user is a prospective student if his wall contains linguistic markers that belong to the spheres of Science or Arts and Humanities. Taking into account that the wall reflects the high school student's self-presentation, we are inclined to think that finding relevant linguistic markers not only shows his/her passion for a particular area of knowledge, but also such soft skills as personal branding, organization, growth mindset, etc.

Our approach consists in classifying user profiles on the basis of presence or absence of keywords manually picked from the communities' total text. The communities were grouped into three categories: humanities, natural science, and physics-mathematics (which corresponds to the most common divisions in Russian universities). It was our decision not to use terminological dictionaries and to extract lexical units by content-analyzing the communities. Their orientation towards the target audience and educational character could provide a better match between the "wall discourse" and "community discourse".

Thematic communities were hand-picked and classified. We sampled 324 popular communities which had from 1 000 to 100 000 participants and dealt with various subject areas: physics and space sciences, mathematics and informatics, engineering, chemistry and materials science, biology and life science, earth science, philosophy and sociology, literary studies, linguistics, psychology, cultural studies, history and law.

We used frequency analysis of community walls (total text amounted to 1 Gb) for creating thematic thesauri. These were used to develop a classifier of the prospective student's leading interests. That is, our task was to identify the preferred subject area of a VKontakte user on the basis of the pre-constructed thesauri. At the first stage of identification we intended to solve a less complex task of identifying and distinguishing between Arts and Humanities students (identified as "humanities" in Table 1) and Science students (identified as "non-humanities" and containing what is normally included in the Life Sciences plus Physics).

Two sets of keywords were created: 432 words in the Arts and Humanities subject area and 246 in the Sciences. These lexical units were chosen by a frequency criterion that was discovered empirically: a word's relative frequency has to be between 0,001% and 0,01%. This range included the words which corresponded to the following two characteristics (here we provide the English equivalents of Russian words found in the total text; all are spelled as one word):

- specialized or terminological nature: amino acid, biodiversity, hominid, ichthyology, geoinformatical, cartographical, space imaging, gerund, unstressed, sketching, allegory, Grand Prince, enthronement, a priori, idealistic, impute, inquest, etc;
- proper names connected with the subject area (Althusser, Wundt, Husserl, Bacon, Herodotus), more common for the Humanities.

Non-specific lexemes and words commonly used in academic writing, Internet slang and jargon that had found their way into the sample were deleted manually. The resulting dictionaries contained lexemes that had low level of polysemy or homonymy, which helped to lower false triggering and made the word vector more precise. At the same time, these lexemes showed enough frequency to be found on the walls of specific users in the form of a repost.

Using the previously constructed dictionaries, we then applied formal quantitative methods of text analysis to the aforementioned samples of computer-mediated discourse to discover the interests of VKontakte users, similar to [13].

As the object of analysis, we used texts on the walls of the first-year Arts and Humanities and Science students at TSU. The texts included reposts, comments and original posts. Minimum amount of data equaled 20 Kb, the number of walls was 50.

The analysis was conducted in the following stages:

1 sampling and normalization;

2 defining the variables for evaluation of objects in the sample, i.e. searching the feature space;

3 determining whether there exist statistically significant differences between the two independent groups for the chosen features;

4 calculation of relative frequency of lexical units.

As a conclusion, we introduce two competing hypotheses:

- H0 – the sampled groups show no significant difference for the chosen feature;
- H1 – the sampled groups differ significantly for the chosen feature.

In the study we used two software products. For text normalization, search for keywords and matrix creation R 3.1.1 programming language was used; for criteria calculation, specialized packages of statistical data processing in STATISTICA 10 were used.

The first stage consisted in normalizing the text. In our study we treat normalization as deletion of punctuation, converting all letters to one case and stemming. These procedures were performed with the R programming language. Stemming was accomplished with the console stemmer Mystem 3.0, developed by Yandex [URL: https://tech.yandex.ru/mystem/].

At the second stage, we searched the users' walls for the words from the thematic dictionaries. With aim of testing the hypotheses we analyzed the walls of first-year students of different faculties. The results of using the relative frequency of words on the students' wall are summarized in Table 1.

Table 1. *Frequency matrix of thematic thesauri on VKontakte users' walls.*

Faculty and VK ID	Thesaurus	
	"Arts & Humanities" Words	"Science and Technical" Words
humanities_id224*****.txt	0,0231	0,0226
humanities_id238*****.txt	0,0208	0,0150
humanities_id330*****.txt	0,0069	0,0188
humanities_id514*****.txt	0,0069	0,0301
humanities_id576*****.txt	0,0463	0,0639
humanities_id902*****.txt	0,0000	0,0000
humanities_id922******.txt	0,0000	0,0188
humanities_id941******.txt	0,0046	0,0000
humanities_id953******.txt	0,0046	0,0226
humanities_id599****.txt	0,0069	0,0150
humanities_id992****.txt	0,0000	0,0000
humanities_id638*****.txt	0,0185	0,0075
humanities_id705*****.txt	0,0000	0,0038
humanities_id825*****.txt	0,1528	0,1015

Faculty and VK ID	Thesaurus	
	"Arts & Humanities" Words	"Science and Technical" Words
non-humanities_id101****.txt	0,0023	0,0075
non-humanities_id109****.txt	0,0069	0,0075
non-humanities_id120****.txt	0,0116	0,0338
non-humanities_id128****.txt	0,0000	0,0000
non-humanities_id128****.txt	0,0093	0,0150
non-humanities_id130****.txt	0,0000	0,0000
non-humanities_id131****.txt	0,0880	0,1429
non-humanities_id137****.txt	0,0046	0,0150
non-humanities_id140****.txt	0,0069	0,0263
non-humanities_id142****.txt	0,0139	0,0075
non-humanities_id145****.txt	0,0162	0,0564
non-humanities_id111****.txt	0,0116	0,0940
non-humanities_id137****.txt	0,0139	0,0038
non-humanities_id138****.txt	0,0023	0,0075

humanities_id902*****.txt	0,0000	0,0451
humanities_id921*****.txt	0,0208	0,0301
humanities_id932*****.txt	0,0208	0,0000
humanitiesos_id823***.txt	0,0116	0,0188
humanitiesos_id8741**.txt	0,0231	0,0376
humanitiesos_id906***.txt	0,0093	0,0263
humanities_id884*****.txt	0,0139	0,0263
humanities_id884*****.txt	0,0139	0,0075
humanities_id896*****.txt	0,0023	0,0113
humanities_id899*****.txt	0,0231	0,0150
humanities_id923****.txt	0,0023	0,0150
humanities_id939****.txt	0,0278	0,0113
humanities_id951****.txt	0,0000	0,0226

non-humanities_id164****.txt	0,0000	0,0038
non-humanities_id172****.txt	0,0023	0,0000
non-humanities_id185****.txt	0,0324	0,0677
non-humanities_id136****.txt	0,0023	0,0301
non-humanities_id157****.txt	0,0000	0,0000
non-humanities_id101****.txt	0,0046	0,0752
non-humanities_id107****.txt	0,0046	0,0000
non-humanities_id138****.txt	0,0278	0,0150

The next stage of analysis was the search for statistically significant differences in independent groups for the features chosen. The sample had a normal distribution (x2 = 3,649, p = 0,056), which permitted us to apply Student's t-test. After studying the results of the t-test we accepted the alternative hypothesis for the "Arts & Humanities" group (p=0,009342) and had to reject it for the other group (see Table 2).

Table 2. *Student's t-test results.*

	t-value	**p**
Students in the Arts & Humanities faculties	-2,713	0,009
Students in the Science and Technical faculties	-1,706	0,094

Hence, the alternative hypothesis H1 is accepted for the "Humanities thesaurus" and rejected for the "non-Humanities" one.

The testing analysis results show that the classification methodology is successful but is not accurate yet. For improving the method, it requires adding two more thesauri on Life Science and Physics-Mathematics (not less than 400 linguistic units) to the Humanities one. We set the objective of achieving a more valid result comparing the adjusted thesaurus with the wall content of 2000 TSU first-year students. It might be interested to compare the frequency of the thematic thesaurus from a first-year student's wall with his/ her academic success. It is going to help testing the hypothesis that an entrant with a high frequency of one thesaurus is more academically successful at the corresponding faculty.

The method of linguistic analysis of the wall content in relation to an entrant's educational interests should go along with the analysis of thematic communities this entrant participates. Entering a community and subscribing to a page in social networks may characterize an entrant's interests. If we choose topics that are relevant to education and cognition out of the entrant's interest spectrum, we might achieve a higher precision of the classification on subject fields.

The study covers the analysis of the thematic content of the communities of 18,000 entrants of the only one town (Tomsk). We have downloaded and generalized the entrants profile data that is connected to the communities, they participate in. From the overall number of the communities we have chosen 959 that have been mentioned in the profiles of 10 or more users. The topics of the

communities have been defined through the expert assessment. Their analysis has been performed manually. As a result, we have made up a classifier of communities and defined a share of every rubric in the total number of communities (Table 3).

Table 3. *Classifier of the thematic communities of the entrants.*

Rubric	Share of the rubric communities in the total number of communities	Rubric	Share of the rubric communities in the total number of communities
entertainment	27,3%	nature and travelling	2,6%
humor	15,2%	interest to a particular university	2,5%
educational	5,9%	sport and health	2,5%
music	5,1%	motivators	2,2%
hobby	5,0%	philosophy and esotery	2,1%
fashion and beauty	4,8%	education	2,0%
art and design	4,2%	languages	1,7%
goods and services	4,2%	literature	1,5%
people, events and firms in the region	3,0%	technics and technologies	1,1%
games and e-sports	2,9%	news, mass media	0,8%
cinema	2,7%	contacts and communication	0,6%

Only some of the rubrics in this classifier have any relation to the entrant's educational interests and can be considered to be markers of inclination to a particular subject field: "art and design", "nature and travelling", "philosophy and esotery", "languages", "literature", "technics and technologies". The share of such communities-markers is only 13,7% out of the total number of communities. Increasing the sample to the number of possible entrants (200,000 people), we suppose to find new thematic rubrics correlating with the additional subject fields. Basing on the present classifier, we plan to calculate for each user a relative frequency of participation in the communities connected to the Humanities, Life Science and Physics-Mathematics. We suppose that uniting these results with the results of the linguistic analysis of the user's wall content is going to give a brighter picture of the entrant's inclination and interest to a particular education program at the university.

4 CONCLUSIONS

Ranging the users within every thematic group (Humanities, Life Science and Physics-Mathematics) according to their level of interest to this or that subject field will make it possible for a university to find the most promising entrants decreasing the number of the target audience from 200 thousand to 10-20 thousand of people and continue individual work with each of the entrants via the social network. This approach that is based on the open big data analysis enables a university with cutting on the marketing expenses and at the same time broadening the geographical coverage, increasing the relevance of the audience, organizing personal communications, leveraging the number of highly motivated entrants and attracting more talented students. Meanwhile this method of revealing the social network users with a particular educational interest may be adapted to the promotion of diverse educational products, such as massive open online courses and vocational education programs.

REFERENCES

[1] Т. В. Алтухова. Социальная компьютерная сеть «ВКонтакте»: жанровая характеристика // Вестник КемГУ 2012 № 4 (52) Т. 3. – С. 21-25.;

[2] Л.И. Ермоленкина, Е.А. Костяшина. Коммуникативно-языковые механизмы формирования этнокультурной идентичности в дискурсивном пространстве интернета // Вестник Томского государственного университета. Культурология и искусствоведение. 2013. №3 (11) – с. 5-15.;

[3] А.В. Щекотуров. Конструирование виртуальной гендерной идентичности подростков на страницах социальной сети «ВКонтакте» // Женщина в российском обществе. 2012. № 4 (65). С. 31-43.

[4] З.И. Резанова. Институциональная и личностная презентация национально-культурной идентичности в интернет-коммуникации: жанровые формы и дискурсивные стратегии // Вестник Томского государственного университета. 2013. № 375. С. 33–41.

[5] А. Е. Войскунский, А. С. Евдокименко, Н. Ю. Федунина. Альтернативная идентичность в социальных сетях // Вестн. Моск. Ун-та. Сер. 14. Психология. 2013. № 1 – С. 66-68.

[6] M.D. Back, J.M., Stopfer, S. Vazire, S. Gaddis, S.C. Schmukle, B. Egloff & S.D. Gosling (2010). Facebook profi les refl ect actual personality not self-idealization. // Psychological Science, 21, 372—374

[7] S.D. Gosling, A.A. Augustine, S. Vazire, N. Holtzman, S. Gaddis (2011). Manifestations of personality in online social networks: self-reported Facebook-related behaviors and observable profile information. // Cyberpsychology, Behavior and Social Networking, 14, 9, 483—488.

[8] Т. В. Алтухова. Электронные и рукописные жанры естественной письменной речи: сопоставительный аспект (на примере граффити и записей на электронной стене) // Вестник КемГУ № 2 (50) 2012. С. 110-116.

[9] А.С. Марковская. Особенности поздравления с днем рождения в социальных сетях // Вестн. Моск. ун-та. Сер. 19. Лингвистика и межкультурная коммуникация. 2013. № 4 – С. 153-159

[10] Е. И. Горошко, Т. Л. Полякова. К построению типологии жанров социальных медий // Жанры речи №2 (12)’2015 – с. 119-127.

[11] Е. И. Горошко. Лингвистика Интернета: формирование дисциплинарной парадигмы / Е.И. Горошко // Жанры и типы текста в научном и медийном дискурсе. – Орел: Картуш, 2007. – Вып. 5. – С. 223-237;

[12] Е.И. Горошкою Современная интернет-коммуникация: структура и основные параметры (коллективная монография) Интернет-коммуникация как новая речевая формация. – М.: Изд-во Наука, Изд-во Флинта, 2012. – 323с. – с.9-52.

[13] N. Mangal, R.Niyogi, A. Milani. Analysis of Users’ Interest Based on Tweets // Computational Science and Its Applications. – Springer, 2016. – V. 9790. – pp. 12-23.

[14] D. Riffe, S. Lacy, F. G. Fico. Analysing media messages: Using quantitative content analysis in research. — Mahwah, NJ : Erlbaum, 2005.