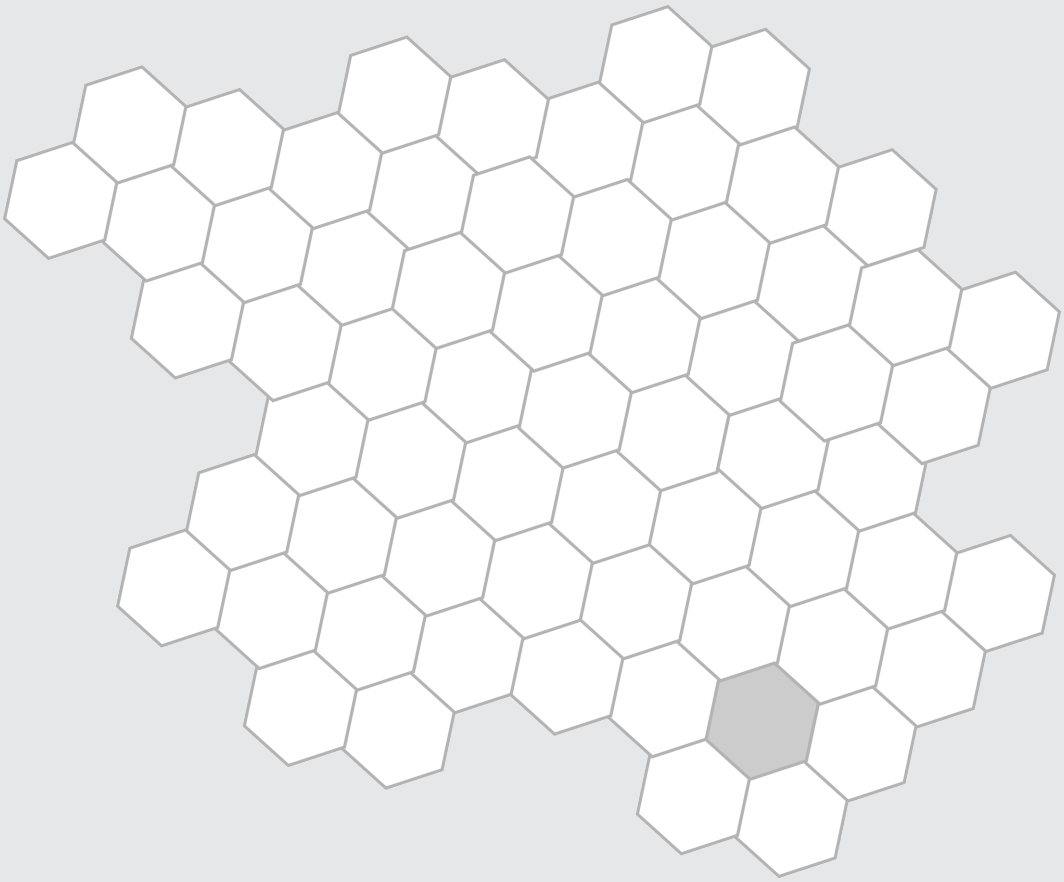
International Journal on

SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining



I, 2010, 1
ISSN 1804-5251

International Journal on SOCIAL MEDIA

MMM: Monitoring, Measurement, and Mining

I, 2010, 1

Editor-in-Chief: Jan Žižka

Publisher's website: www.konvoj.cz
E-mail: konvoj@konvoj.cz, SoNet.RC@gmail.com
ISSN 1804-5251

Journal Profile

International Journal on Social Media MMM: Monitoring, Measurement, and Mining is an international scientific and refereed journal focused on questions and progress of social media, especially on their monitoring, measurement, analysis, and mining in social networks, e.g. Sentiment/Opinion Analysis in Natural-Language Text Documents, Algorithms, Methods, and Technologies for Building and Analysing Social Networks, Applications in the Area of Social Activities, Knowledge Mining and Discovery in Natural Languages Used in Social Networks, Medical, Economic, and Environmental Applications in Social Networks, etc.

International Journal on Social Media MMM: Monitoring, Measurement, and Mining seeks to share new knowledge, processes and methods. The journal publishes original works, project solutions, case studies, reviews and educational papers. Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere.

Two issues per year provide a forum for distinguished as well as young authors. A shortened thesis as well as final reports of projects supported by grant agencies are accepted for publishing.

All papers are refereed through a peer review process.

Submissions should be send in the PDF form via email to the following address: `SoNet.RC@gmail.com`.

Accepted papers are to be prepared according to the instructions available at `http://www.konvoj.cz/journals/mmm/`.

International Journal on Social Media MMM:
Monitoring, Measurement, and Mining,
I, 2010, 1
Published by KONVOJ, spol. s r. o. (ID CZ 47915391),
Bystřinova 4, CZ 612 00 Brno

The journal comes out twice a year. Printed in the Czech Republic.

Registration number of Ministry of Culture: MK ČR 19640
Issued in Brno, September 3, 2010.

ISSN 1804-5251

Contents

Editorial

We would like to introduce here the first and, at the same time, special issue of our new international journal. As a convenient opportunity, we take the advantage of organising the second international workshop SoNet-2010 Social Networks: Computing and Mining, September 3–5, 2010, Mendel University Brno, Czech Republic. This small workshop, aiming at selected branches from the particularly topical problem area connected with social networks, brings articles on an array of topics that are related to computing and data/knowledge-mining. The defined workshop intent included particularly the following areas:

- Algorithms, methods, and technologies for building and analysing social networks.
- Automatic analysis of sentiment, subjectivity, and opinions in social networks.
- Knowledge mining and discovery in natural languages used in social networks.
- Social networks and Web 2.0/3.0.
- Economic, business, environmental, and medical applications of social networks.
- Applications in the area of social activities.

This journal issue contains a collection of articles carefully selected from submitted research results and intentions. In addition, SoNet-2010 could welcome two invited outstanding scientists: Prof. Michael Thelwall from the Wolverhampton University, United Kingdom, and Dr. Alexander Troussov from the IBM Dublin Center for Advanced Studies, Ireland. Thus, the workshop participants could listen to up-to-day and very interesting talks and presentations, which have been included in this special journal issue.

The SoNet-2010 international workshop was organised by people from the Department of Informatics, Mendel University Brno, Czech Republic and fLexSem Laboratory, Department of French and Romance Philology, Autonomous University of Barcelona, Spain. The event took place at the Faculty of Business and Economics, Mendel University Brno. Thanks to the support provided by the Faculty and University, the Spanish researchers from Barcelona, and by the Konvoj Publishing Company, the event passed off smoothly. As one of its results, SoNet-2010 has contributed to broadening the already existing research cooperation as well as creating new promising scientific collaborations in the specific areas.

Here, we would also like to thank to all SoNet-2010 Organising and Program Committee members for their conscientious and time-consuming work. Without their effort, the workshop could not be so successful.

We hope and believe that the workshop results published in this journal issue will attract attention of new and more scientists and researchers.

On behalf of all people participating in the SoNet-2010 realisation,

Jan Žižka
Department of Informatics/SoNet Research Center
Faculty of Business and Economics
Mendel University Brno, Czech Republic

SoNet 2010

Programme Committee

Jan Žižka, Co-Chair (Mendel University, Czech Republic)
Xavier Blanco, Co-Chair (Autonomous University of Barcelona, Spain)
Angels Catena (Autonomous University of Barcelona, Spain)
František Dařena (Mendel University, Czech Republic)
Jiří Hřebíček (Masaryk Univesity, Czech Republic)
Tomáš Hudík (Moravia Worldwide, Brno, Czech Republic)
Pavel Makogonov (Mixtec Technological University, Mexico)
Natalia Ponomareva (Wolverhampton University, United Kingdom)
Paolo Rosso (Technical University of Valencia, Spain)
Michael Štencl (Mendel University, Czech Republic)
Michael Thelwall (Wolverhampton University, United Kingdom)
Alexander Troussov (IBM Dublin, Ireland)
Victor Zakharov (St. Petersburg State University, Russia)

Organising Committee

Oldřich Trenz, Co-Chair (Mendel University, Czech Republic)
Mikhail Alexandrov, Co-Chair (Autonomous University of Barcelona, Spain)
Lorraine Baqué (Autonomous University of Barcelona, Spain)
Tomáš Hála (Mendel University and Konvoj Publishing Company, Czech Rep.)
Jaromír Landa (Mendel University, Czech Republic)
David Procházka (Mendel University, Czech Republic)
Arnošt Svoboda (Masaryk University, Czech Republic)

Schedule

Friday, September 3, 2010

09:00–10:00 … Registration

10:00–10:15 … Welcome, Opening

10:15–11:00 Michael Thelwall:
Sentiment Strength Detection for Social Network Site Comments (invited talk)

11:00–11:15 … Coffee Break

11:15–11:40 Olga Kaurova, Mikhail Alexandrov, Natalia Ponomareva:
The Study of Sentiment Word Granularity for Opnion Analysis

11:40–12:05 Joseph Zernik:
Data Mining as a Civic Duty: Online Public Prisoners Registration Systems

12:05–13:30 … Lunch Break

13:30–14:15 Alexander Troussov:
Harnessing the Power of Social Context – 1. (invited talk)

14:15–14:40 Martin Adámek:
CB-Radio in Road Traffic as Social Network and Information Technology

14:40–15:00 … Coffee Break

15:00–15:25 Olga Ogurtsova, Mikhail Alexandrov, Xavier Blanco:
A Look on Wikipedia Readibility: Language, Domain and Style

15:25–15:50 Gabriel Lukáč:
A Proposal of an Approach to Extracting Conceptual Descriptions of Hyper-linked Text Documents

15:50–16:15 David Sousedík, Ladislav Buřita:
Social Network as a Part of the Interactive Environment for Starting Enterpreneurs

Saturday, September 4, 2010

08:30–09:00 … Registration

09:00–09:45 Alexander Troussov:
Harnessing the Power of Social Context – 2. (invited talk)

09:45–10:10 Joseph Zernik:
Data Mining of Online Judicial Records of the Networked US Federal Courts

10:10–10:30 … Coffee Break

10:30–10:55 Ronan McHugh:
Persuading Collaboration: Analysing Persuasion in Online Collaboration Sites

10:55–11:20 Angels Catena, Mikhail Alexandrov, Natalia Ponomareva:
Opinion Analysis of Publications on Economics with Limited Sentiment Vocabulary

11:20–13:00 … Lunch Break

13:00–18:00 … Brno Sightseeing Tour (for interested participants – with a guide, free of charge)

Sunday, September 5, 2010

09:00–09:45 Michael Thelwall:
Analysing the Social Context of Social Network Sites: The Case of MySpace
(invited talk)

09:45–10:10 Eva Eckenhofer:
Do Central Players Perform Better?

10:10–10:35 Jan Panuš:
Modified Local search Algorithm as a Tool for Analysing Social Networks

10:35–11:00 Samoylenko O. M., Ilona Batsurovska, Ruchinska N. S.:
Methology of Using WEB 2.0 Services for Higher Education

10:50–12:30 … Concluding, Discussing Future Cooperations

Invited lectures (abstracts)

Sentiment Strength Detection for Social Network Site Comments

Michael Thelwall
School of Computing & IT, University of Wolverhampton, United Kingdom
e-mail: M.Thelwall@wlv.ac.uk

Abstract

This talk describes SentiStrength, a system to detect the strength of positive and negative sentiment expressed in short informal text, such as in the public comments exchanged between friends in social network sites. SentiStrength uses a list of sentiment-bearing words annotated with polarity and strength as the core of its algorithm. Scores based on these words are then enhanced with a variety of features gained from shallow parsing of the text. Some of these features are generic, such as negating terms and booster words, whereas others are specific to informal text, such as emoticons and repeated characters as a device to convey enhanced emotion. Experiments with SentiStrength applied to public comments from the popular social network site MySpace shows that it outperforms a range of machine learning methods for positive sentiment but not for negative sentiment.

Analysing the Social Context of Social Network Sites: The Case of MySpace

Michael Thelwall

School of Computing & IT, University of Wolverhampton, United Kingdom
e-mail: M.Thelwall@wlv.ac.uk

Abstract

The social network site MySpace, once the most popular web site for US Internet users and still a popular site, is oriented to both music and youth. From a research perspective, its relatively open nature makes it easy to conduct large-scale research on its members. This talk illustrates the potential for social network research using computing techniques by reporting the results of a large-scale study of social issues related to MySpace members. The findings include the impact of gender on communication, the age profile of members and the types of information revealed by members in their public profiles.

Biographical note

Michael Thelwall, Professor of Information Science, Webometrics and cybermetrics researcher: Developing quantitative methods for Internet phenomena, including hyperlinks and Web 2.0 social networks. Micheal Thellwal is an author and co-author of more than 150 refereed journal articles, five book chapters, two encyclopedia articles, and many confeence research papers. He is also an author of two books: Introduction to webometrics: Quantitative web research for the social sciences, and Link analysis: An information science approach. For more details, see his web page (http://www.scit.wlv.ac.uk/~cm1993/mycv.html)

Harnessing the Power of Social Context

Alexander Troussov
IBM Dublin, Ireland
e-mail: atrousso@ie.ibm.com

Abstract

We live in an increasingly interconnected world of socio-technological systems, in which technological infrastructures composed of many layers are interoperating within a social context that drives their everyday use and development. Nowadays, most of the digital content is generated within public systems like Facebook, Delicious, Twitter, blog and wiki systems. These applications have transformed the Web from a mere document collection into a highly interconnected social space where documents are actively exchanged, filtered, organized, discussed and edited collaboratively. The emergence of the Social Web opens up unforeseen opportunities for observing social behaviour by tracing social interaction on the Web. In these socio-technological systems 'everything is deeply intertwingled' using the term coined by the pioneer of the information technologies Ted Nelson: people are connected to other people and to 'non-human agents' such as documents, datasets, analytic tools, tags and concepts. These networks become increasingly multidimensional providing rich context for network mining and understanding the role of particular nodes representing both people and digital content. In this talk we show how to represent to our formal reasoning and to model social context as knowledge using network models to aggregate heterogeneous information. We show how the social context could be efficiently used for well understood tasks of natural language processing as well as for novel applications such as social recommender systems which aim to alleviate information overload for social media users by presenting the most attractive and relevant content.

Biographical note

Alexander Troussov, Ph.D., is chief scientist at the IBM Dublin Centre for Advanced Studies and chief scientist of the IBM LanguageWare group. He is a joint author of more than 30 peer-reviewed research publications and has five patents pending. As the LanguageWare group architect, Dr. Troussov developed new methods for the optimisation of finite state processing for morphological analysis, worked on computational models for compounding languages, and worked on other problems of subsententional text processing. As CAS chief scientist, he leads IBM's participation in the NEPOMUK project. More information is available at the IBM web site (http://www.alphaworks.ibm.com/tech/galaxy, https://www-927.ibm.com/ibm/cas/sites/dublin/).

Full papers

CB Radio in Road Traffic As Social Network and Information Technology

Martin Adámek

Department of Information Technologies, Faculty of Informatics and Management, University of Hradec Králové, Rokitanského 62, 500 03 Hradec Králové, Czech Republic
e-mail: martin.adamek@uhk.cz

Abstract

The aim of this paper is to introduce CB radio, its possibilities when using it as a source of traffic information, and author's empirical experience with long term practical using it. Also to describe a gross draft of plan of research the aim of which will be to check and evidence if CB radio is better than other channels for distributing and gaining traffic information.

Results of the future research can be useful for the next development in traffic information technologies.

Key words

traffic information; social network as information technology; CB radio; CB radio; citizen band; radio; transceivers; road traffic

Biographical note

Martin Adámek is a PhD student who used to work as van driver during his master's studies.

User of CB radio on Czech and Slovak standard since early 1990's and user of CB radio on Polish standard since February 2010. Having two transceivers on the dashboard and three long antennas (one for receiving of common radio broadcasting) on the roof of his car to have as good traffic information as possible.

An Introduction – Entering information about CB radio

Various information technologies are being used as source of traffic information for drivers. Besides the well-known technologies like radio broadcasting with RDS or GPS navigation systems with RDS-TMC, it is CB radio system of transceivers too.

This paper is based on author's empirical long-term (ca. 15–17 years) personal experience with using of CB radio, mainly in road traffic.

CB radio (CB = citizen band, civil band) is standard of transceivers whose using of which is allowed generally for public. Standards are not the same in all countries but in each country a channel (frequency) exists which is reserved for communication between drivers and for sharing traffic information. CB uses radio band around 27 MHz (wave length around 11 metres).

Drivers of many trucks, some vans, few off-road expedition cars, and just rarely of common passenger cars use CB in the Czech Republic. So, CB radio is absolutely unknown technology for the most of Czech people. Approximately the same situation as in the Czech Republic is in Slovakia, whereas CB is installed in many cars, including passenger ones in Poland. So CB radio is well known there. There are three reasons for this big difference between the two neighbour countries:

- Polish police is stricter and has bigger respect from its local citizens than Czech police.
- Polish people have wilder character, Polish drivers are faster and more risking.
- Polish AM CB transceivers are ca. 2 to 3 times cheaper than Czech FM ones.

Comparison of Czech and Polish environment is based on author's knowledge of Poland, on whole-life living on state border with Poland, on studying Polish language and culture, on travelling through Poland, and on this year's one semester study stay in Poland.

Truck drivers in many countries, including non-European, use CB radio – but author of this paper has personal experience with Czech, Slovak and with Polish CB radio only.

Brief history of using of CB radio in the Czech Republic

The way of using CB radio has changed with the time progress.

It was available since the beginning of 1990's of 20th century in Czechoslovakia, when GSM phones were not available, price of NMT phone corresponded to ca. 6 month salaries (charges were high too), and waiting period for a fixed phone line could be up to 3 years, sometimes even with shared line.

So CB radio was being used as communication technology for families, companies, organisations or friends. Emergency channel used to be monitored by EMS, police, or metro police in many Czech towns in this time. It was possible to call EMS to traffic accident by CB (verified).

Whereas today, at the beginning of the 21st century, it is possible to get functioning cell phone free of charge; charges are quite low, modern transceivers on new standard PMR (Personal Mobile Radio; band around 446 MHz; lower price, size, weight, battery consumption, and range) are available for sport, outdoor activities and other events. And many internet services exist for unimportant social cost-free conversation (chatting) between people.

So CB radio is being used already almost by drivers only today. Although it is quite an old technology, it still has its place on dashboards of trucks, in addition to the most modern GPS navigation systems equipped with RDS-TMC input and broadcast radio receivers with RDS.

Reason of persisting popularity of CB between truck drivers is relevance of traffic information.

Technical principles and law conditions

CB radio is being used without repeaters on traffic channels. It means direct transmitting from transceiver to receiver and limited range. Between two cars, it can be 100 m in direct visibility – with broken antenna during high radio traffic; as well as it can be 40 km passing a hill – when good antenna and uncertified amplifier (making power e.g. 100 W instead of allowed 4 W) is used.

Transmission range depends mostly on the position and equipment of transmitting side. So, communication paths (relations) can be directed (oriented) in network of mass CB communication.

A network member has connection with two other members who are not connected mutually. Each state of network and paths is very temporary, it is valid for a definite instant only because summation of speeds of two cars in opposite direction can be ca. 3 km per minute, so conditions are changing quite fast during conversation when cars are going near or departing.

CB transceivers are half duplex. It means that each TransCeiver can be only transmitting or only receiving at any moment. Maximally one station can be transmitting on one channel in one geographic location (range of transmitting station). All other stations on the same channel in the range can be listening. When two or more stations are transmitting too near on the same channel, signals jam and some drivers cannot hear anything while some can hear just the stronger (nearer) station. Group communication ('conference', by phones vocabulary) is big advantage of radios and it is a basic precondition for using radios as social network and information technology. It is not necessary to establish connection between two or more concrete users. When a channel is empty anyone can start transmitting; and everyone who is in the range can hear him immediately. 'Conference' is 'established' immediately without any complicated procedure (dialling, ringing or technical joining), so everyone can add his piece of information or opinion almost immediately. This is the strength of CB in its role of social network and IT.

Various national standards exist. Some stations can just jam each other and not to communicate.

Empirical experience with CB in road traffic

Social network

CB radio, either from the point of view as traffic information technology or as system for group communication is a network of people equipped by transceivers.

Usability of CB radio system depends on people. It is possible to see especially in the Czech and Slovak Republics, where it is being used almost in trucks and few vans only, that it does not operate during weekends when truck driving is prohibited, and on side roads where truck traffic is not present. This can be better in Poland where many passenger cars are equipped with CB (but some level of road traffic is still necessary to cause enough volume of radio traffic).

New term 'social network' has been used quite often in ICT world during the few last years. Although this term is much newer than CB radio it is possible to classify soft system of CB radio and its users as social network (network based and depending on users, on people).

CB has the same problems as other social networks when being used as information technology – human element can bring noise (by unrelated disturbing nonsense communication), false information (by mistake, or misunderstanding) or quite impolite expressions sometimes, too.

Information technology

While common radio broadcasting distribute information about traffic in the whole country or region,

CB radio has limited range (from hundreds of meters to several kilometres). So users are not being disturbed by non-interesting 'information' and can work with real useful information, related to area or road where they are.

Each piece of traffic information has to wait for its dedicated time in broadcast time schedule. So traffic information has big delay in broadcasting. Additionally, many radio stations exist. Whereas each piece of information is distributed immediately at CB band – when adriver sees a radar, accident or another danger, he tells it to other drivers, so delay of transfer of information is few seconds only at CB band, instead of tens of minutes at radio broadcasting.

So CB radio is instant and local in comparison with radio broadcasting and provides much more relevant information.

Additionally, it is possible to ask other drivers for some specific information, needed at a given time (traffic situation in front of the driver, navigation to concrete destination point, petrol station or toll sale, or other local problems).

Each piece of information is turning around in circle in some limited (thus related) area around the place of occurrence of an event. It is being constantly repeated and continuously actualised during the whole period of event validity by new drivers who are coming to that place from all directions.

Because of diversity in the range of transmitting and positions of drivers, a piece of information is forwarded by some driver on demand of an other driver who has heard that the first driver is thanking for receiving a piece of information.

Everything aforesaid ensues from long-term personal empirical experience. Author of this paper has clear opinion about usefulness of CB radio in a car (that is why he uses it – as well as most of truck drivers) but it is necessary to support the hypothesis by hard numbers.

A research about this topic already exists. Research was done by asking 1,200 transport companies in California, USA by Regan and Golob (1999). It shows the same experience – *CB radio is the best way of gaining traffic information for drivers*,

The strength of social network (not depending on used communication technology) as source of traffic information is also confirmed by the same paper from Regan and Golob (1999) – *the best source of traffic information for dispatchers are reports from their drivers on the road*.

Gross concept of plan of research

The aim of this research will be to explore, verify and evidence possibilities and benefits of CB radio as a source of traffic information.

First preliminary research has been already done as part of preparation of this paper during the last months. It has not involved a large statistical sample. The main aim of this research was to try punctual logging of traffic information into the protocol to verify or change the scheme of the form.

Main research

The topic of main research will be to watch float of traffic information on various information channels (sources) simultaneously with the aim to verify speed, quality

and geographic relevance of distributing traffic information by this information channels:

- CB radio (Czech and Slovak traffic channel; Original idea to monitor more national channels was rejected because of too low foreign radio traffic)
- Public radio broadcasting (one full area and one regional station; no information about radars)
- Commercial radio broadcasting (ca. two full area stations and ca. three regional stations)
- NDIC website (official National traffic information centre, source for public radio broadcasting)
- SMS system `radary.cz`

Monitoring can be being done:

- In interesting places (traffic knots; main roads during some traffic event)
- By monitoring of frequency of repeating of the same information
- In various weekdays, in various time, on roads of various size (importance)

Another researches

- Central information service operated in traffic knot (logging of information float into a protocol)
- Watching how many drivers pay fine for speeding have CB antenna
- Watching distance-light signals and gesticulation of oncoming drivers
- Counting vehicles equipped and unequipped by CB antenna, with information about country and category of the vehicle; verifying how many drivers with CB antenna use CB
- Everything for various types (sizes) of roads, various weekdays, various time
- Watching traffic information and comparing it with actual state while driving

Research should show and evidence if CB radio is really as useful as it seems based on empirical experience after ca. 15 years of occasional using it.

Conclusion

CB radio provides much more related information (in aspect of location and time) than radio broadcasting, because it is instant (few seconds) and local (several kilometres). While radio broadcasting has unacceptable time delay (tens of minutes) and place irrelevance (hundreds of kilometres). And information integrity (non-absence of information) is better on CB, too.

But usability of CB radio depends profoundly on type of road and on weekday + time because it depends on truck traffic.

It is necessary to verify empirically drawn conclusions by statistically acceptable way.

Foreign drivers usually respect prohibition of using their radios, so it is not possible to monitor information float on various national CB traffic channels. But

it is possible to compare information float on Czech CB traffic channel with other sources of traffic information.

Acceptable results could serve as inspiration for developers of sophisticated navigation and traffic information systems. It could also prove that people (users) are still quite an important part of the world, and that new traffic information systems could be duplex and could involve users as direct and fast source of information.

References

Regan, A. C., Golob, T. F. (July 1999): *Freight Operators´Perceptions of Congestion Problems and the Application of Advanced Technologies: Results from a 1998 Survey of 1200 Companies Operating in California* [on-line]. [Irvine (U.S.A.)] : University of California, [cit. 2010-07-19]. Available at: `http://128.200.36.2/its/publications/papers/CLIFS/UCI-ITS-LI-WP-99-7.pdf`.

Mika, P. (December 2006): *Social Networks and the Semantic Web* [on-line]. [Amsterdam, (Netherlands)] : Vrije Universiteit, [cit. 2009-10-03]. Available at: `http://dare.ubvu.vu.nl/bitstream/1871/13263/5/7915.pdf`.

Opinion Analysis of Publications on Economics with a Limited Vocabulary of Sentiments

Angels Catena
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: angels.catena@uab.cat

Mikhail Alexandrov
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: malexandrov@mail.ru

Natalia Ponomareva
Statistical Cybermetrics Research Group, School of Computing and IT, University of Wolverhampton, Stafford Str, WV1 1SB Wolverhampton, UK
e-mail: nata.ponomareva@wlv.ac.uk

Abstract

Opinion Analysis (OA) is a part of so-called Sentiment/Subjectivity Analysis, which aims to evaluate the author's personal characteristics and his/her attitude to objects and events. The existing well-known OA-systems use large vocabularies of classified sentiments (thousands of words) to give a positive or negative answer1. In the paper we consider another case when the sentiment vocabulary is very limited (one-two hundreds of words) and the answer list includes an additional neutral category. We study OA-accuracy of Spanish documents related to economic crisis. Decision-making is implemented on regression model trained on examples. We show the dependency of OA-quality on a) granularity of sentiments and opinions b) rules used in regression model. We also compare the results with those obtained in Bo Pang and Maite Taboada research groups. In case of binary classification of sentiments and opinions the results prove to be similar.

Key words opinion analysis; sentiment vocabulary; regression model

Biographical note

Angels Catena is a member of the fLexSem Research group of the Autonomous University of Barcelona (UAB) in Spain. She works as assistant professor for the Department of French and Romance Philology (UAB). She is a linguist and an author of several publications related to lexicography, translation studies and pedagogical lexicography. Her current topics of research are paraphrase recognition and social networks.

Mikhail Alexandrov is a member of the fLexSem Research group at the Autonomous University of Barcelona in Spain. He is a professor of the Academy of National Economy under Russian Government. He is an applied mathematician and author of numerous publications related to mathematical modelling and natural language processing. His current topics of research are machine learning (inductive modelling, clustering) and internet technologies (social networking).

Natalia Ponomareva is a PhD student at the University of Wolverhampton. She received her master's degree from the Technical University of Valencia. Her research interests include Sentiment, Sentiment Transfer and Machine Learning for NLP. She is author of more than 10 scientific publications in international conferences and journals.

Introduction

Paper terminology

In this paper we use the following terminology:

Sentiments are words having a positive or negative sense in Opinion Analysis (OA). Sentiments are presented in the form of 4 vocabularies: nouns, verbs, adjectives and adverbs

Sentiment classification is a list of sentiment categories. We use two classifications: a rough two-level classification (positive and negative) and a detailed classification of 4 levels (very positive, positive, very negative and negative). Individual sentiment contribution in the first case is equal to 1 and in the second case to 1 and 0.5 respectively.

Opinion classification is a list of opinion categories. We use two classifications: a rough two-level classification (positive and negative) and a detailed classification of 4 levels (very positive, positive, very negative and negative). When the neutral category is used, the rough classification includes 3 categories (positive, negative and neutral) and the detailed classification includes 5 categories (very positive, positive, neutral, negative and very negative). According to these categories an expert evaluates each document using scales (-1,1) and (-1,0,1) for the rough classification, and (-1,-0.5,0.5,1) and (-1,-0.5,0,0.5,1) for the detailed classification.

Models of OA are all combinations of sentiment classifications with opinion classifications. It is easy to see that we have 4 such combinations: the rough categories of sentiments and the rough categories of opinions, the detailed categories of sentiments and the rough categories of opinions, etc.

The **regression model** for OA is lineal equation, the value of which is transformed into one of the categories from the opinion classifications. Arguments of the regression are so-called linguistic variables. Such a model is trained on examples prepared by experts, and then it is used on new texts. By 'lineal' we mean: a) linearity with respect to coefficients; b) linearity with respect to linguistic variables.

Linguistic variables can reflect the contribution of all positive sentiments, all negative sentiments, or their total contribution (the sum). When we have separate linguistic variables for positive and negative sentiments we deal with two-parameter regression. When the linguistic variable is a composition of positive and negative sentiments (i.e. the sum) we deal with one-parameter regression.

The **models for decision-making** on regression are the set of regression values and correspondent categories of opinions. One can change these rules and obtain different results.

Related works and problem settings

The general approach to OA that we follow in this paper is Machine Learning. Such an approach was proposed and developed by Pang et al. (2002) and Pang

and Lee (2004, 2008). Pang and Lee (2002) considered the domain of movie reviews. Their data included positive, negative and neutral reviews, but the authors concentrated only on positive and negative ones (700 and 700). They experimented with three standard methods: Naive Bayes classifier, maximum entropy classifier, and support vector machines with different sets of sentiments (2,600–32,300).

Maite Taboada's semantic orientation calculator SO-Calc is a well-known OA-system (Taboada et al., 2006; Brooke et al., 2009). SO-Calc uses 4 open vocabularies (nouns, adjectives, adverbs and verbs, about 5,000 sentiments in total). All sentiments are ranged on a 10-point scale. SO-Calc uses a regression model for binary classification. The authors exprimented with a set of positive and negative reviews (200+200) covering 8 topics: books, cars, movies, etc.

Our tools are similar to those of Maite Taboada's group. We use 4 vocabularies, a program for calculation of sentiment contributions to a given document, and a regression model for decision-making. The difference consists in the size and granularity of vocabularies, granularity of opinions, and in flexible rules for decision-making on the regression equation.

The subject under consideration is documents related to the economic crisis: interviews, surveys, analytical papers, etc. We consider the following problems:

1) Sensibility of results to sentiment classification
2) Sensibility of results to opinion classification
3) Sensibility of results to rules of decision-making on regression

The paper consists of 5 sections. Section 2 describes data under consideration and sentiment vocabularies. Section 3 shows how the regression model is constructed and evaluated. Section 4 presents the results of all experiments. Section 5 contains the discussion of results and proposals for future work.

Parameterisation

Documents

The initial material consists of 50 papers in Spanish and Catalan. The papers vary greatly in length: from one to several pages. All papers were evaluated by two experts using a 4-point scale. Table 1 contains the titles (in English) and points of these documents. Table 2 shows the distribution of papers on categories. It is easy to see that the neutral category is small enough in comparison with polar categories in the rough opinion classification.

With the rough opinion classification all points 0.5 and -0.5 are transformed into 1 and -1 respectively.

Table 1: List of documents with their points (part of full list)

No	Title	Points
1	BBVA thinks that Spanish economy has already passed the point of recession	0.5
2	Spanish economy could enter to a long phase of stagnation according IESE	-1
3	The great problem of Spanish economy	-0.5
4	French economy grows in 3rd semester	0

In the paper we consider all possible models of OA. They are described in Table 3.

Vocabularies

The linguistic resources for OA are presented in the form of 4 vocabularies: nouns, verbs, adjectives and adverbs. All sentiments were ranged on a 4-point scale. Table 4 contains vocabulary descriptions. Table 5 presents the vocabulary of adverbs without English translation.

With the rough sentiment classification all points 0.5 and -0.5 are transformed into 1 and -1 respectively.

Parameterised documents

Document parameterisation consists of two steps:

- Evaluation of the sentiment contribution to a given document;
- Formation of the value for a linguistic variable(s).

In order to complete the first step we developed a program on Python. The input data of this program are the 4 vocabularies described above. The output data are numbers of positive and negative sentiments and their summary contribution

Table 2: Distribution of papers on categories

No	Category	Number	% with neutral categ.	% without neutral categ.
1	Very positive	3	6	7
2	Positive	17	34	39
3	Neutral	7	14	
4	Negative	15	30	35
5	Very negative	8	16	19

Table 3: Models of OA considered in the paper

No	Sentiment classification	Opinion classification
1	rough (2 categories)	rough (2 or 3 categories)
2	detailed (4 categories)	rough (2 or 3 categories)
3	rough (2 categories)	detailed (4 or 5 categories)
4	detailed (4 categories)	detailed (4 or 5 categories)

Table 4: Components of vocabularies

No.	Vocabulary	Size	Very positive	Positive	Very negative	Negative
1	Nouns	58	9	16	16	17
2	Verbs	50	8	15	13	14
3	Adjectives	47	9	9	15	14
4	Adverbs	33	8	9	7	9
	Total	188	34	49	51	54

to a given document. The program calculates the contribution simultaneously for the rough and the detailed sentiment classifications. Obviously, in the case of the rough classification the number of positive and negative sentiments coincides with their contribution (in absolute value) because each sentiment has a weight equal to 1. Table 6 shows a part of the full table prepared by the Python program. The contributions are located in the last four columns.

In order to complete the second step let us have a look at the distribution of positive and negative contributions within the framework of the detailed classification. Figure 1 shows this distribution. One sees a large spread. Since the regression value changes in a limited interval [-1,1] such a large spread of contributions can lead to inconsistent regression.

To exclude this effect we normalise all contributions on the total number of sentiments. Therefore we have:
P_L = Positive contribution / Total number of sentiments
N_L = Negative contribution / Total number of sentiments
where P_L and N_L stand for positive and negative linguistic variables respectively.

Both linguistic variables now are located within the interval [-1,1]. Figure 2 shows their joint distribution.

We calculated the coefficient of correlation between P_L and N_L. It proved to be 0.91. With such a correlation we construct one linguistic variable instead of two: $L = P_L + N_L$

The set of linguistic variables is the final result of document parameterisation.

The regression model

Rules for decision-making on regression

Our goal is to construct and to test regression equations for all 8 models of OA reflected in Table 3. The regression equation is presented in the form:

$$R = a + bL$$

Here: R is the value of regression equation, L is the linguistic variable and a and b are unknown coefficients. Models for decision-making on regression are presented in Tables 7–10.

Table 5: Vocabulary of adverbs (in Spanish)

Positivo (0.5)	Muy positivo (1)	Negativo (-0.5)	Muy negativo (-1)
delante	positivamente	debajo	negativamente
suavemente	estupendamente	demasiado	duramente
favorablemente	brillantemente	inevitablemente	gravemente
tímidamente	excelentemente	difícilmente	cruelmente
moderamente	felizmente	insuficientemente	alarmante
suficientemente	bien	tardíamente	desgraciadamente
oportunamente	gracias (a)	seriamente	dramáticamente
adecuadamente	acertadamente	lentamente	
convenientemente		precipitadamente	

Table 6: The results of Python program (part of full table)

Points	Text	Number of pos. sentim.	Number of neg. sentim.	Detailed classif., Positive	Detailed classif., Negative	Rough classif., Positive	Rough classif., Negative
0.5	T1.txt	19	11	9.5	-6.5	19	-11
-1	T2.txt	15	39	7.5	-27.5	15	-39
-0.5	T3.txt	2	6	1	-4	2	-6
-1	T4.txt	8	11	5	-8	8	-11
0.5	T5.txt	11	8	6	-4.5	11	-8
-0.5	T6.txt	22	14	11.5	-8.5	22	-14
-0.5	T7.txt	28	42	16	-32	28	-42

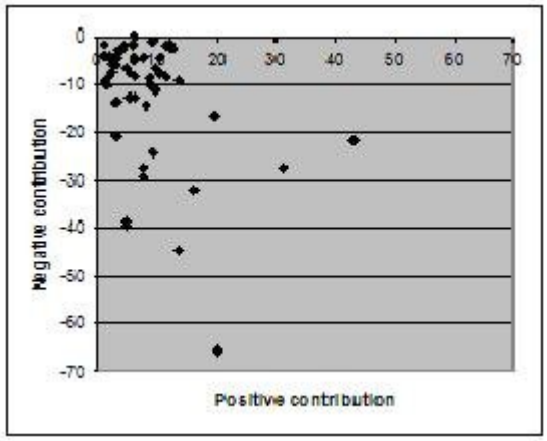


Figure 1:
Distribution before normalisation

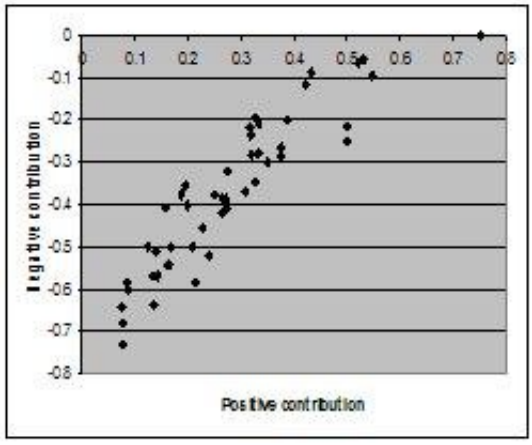


Figure 2:
Distribution after normalisation

Table 7: Rules for decision-making with 2 categories of opinions (the neutral category is absent)

Regression value	Opinion
$R < 0$	Negative
$R \geq 0$	Positive

Table 8: Rules for decision-making with 3 categories of opinions (the neutral category is present)

Regression value	Opinion
$R < -0.5$	Negative
$0.5 \leq R \leq 0.5$	Neutral
$R \geq 0$	Positive

Table 9: Rules for decision-making with 4 categories of opinions (neutral category is absent)

Regression value	Opinion
$R < -0.75$	Very negative
$-0.75 \leq R < 0$	Negative
$0 \leq R \leq 0.75$	Positive
$R > 0.75$	Very positive

Table 10: Rules for decision-making with 5 categories of opinions (the neutral category is present)

Regression value	Opinion
$R < -0.75$	Very negative
$-0.75 \leq R < -0.25$	Negative
$-0.25 \leq R \leq 0.25$	Neutral
$0.25 < R \leq 0.75$	Positive
$R > 0.75$	Very positive

These intervals in Tables 7–10 should be assigned according to prior information about the distribution of opinion categories on axis R, but initially the interval stated above seems to be the most natural.

Evaluation of model quality

To evaluate the model quality we should select an index or indexes reflecting this quality. It can be the accuracy for all categories or for each category, the so-called F-measure for all categories or for each category, etc. A good survey of indexes for problems of classification is presented by Pinto (2008). In this paper we use the total accuracy, which is measured by the simplest formula:

$$\text{accuracy} = N_c/N_e$$

Here: N_c is the number of coincidences between expert opinions and model replies, N_e is the total number of experiments. N_e=50 when we use the neutral category and N_e=43 when we do not use it. One should say that the accuracy cannot be considered as the final quality index. It is necessary to take into account a so-called *Baseline*. It is the lowest value of accuracy which can be obtained on a given data set. Usually the Baseline is equal to the probability of the most frequent category. For this reason we introduce a so-called *adjusted accuracy*:

$$\text{adjusted accuracy} = \text{accuracy} - \text{Baseline}.$$

We can calculate the Baseline for all OA models using Table 3. The results are presented in Table 11.

To evaluate the *accuracy* we use the standard procedure of cross-validation. In this procedure all data are divided into a training and a control set. Then the model constructed on the training set is tested on the control one. Such an experiment is repeated on several partitions and the average error is calculated. In this paper we use leave-one-out cross validation when the training set contains Ne-1 data and the control set contains one data. Cross-validation is performed with a well-known package Weka (Weka, 2009).

Experiments

Evaluation of Opinion Analysis without the neutral category

In this series of experiments we studied the accuracy of decision-making on a set of 43 documents. Table 12 contains the results of the calculation. These results are presented in graphic form in Figure 3.

Table 11: Baselines for different OA models

Opinion classification	Number of categories	Baseline
Rough classification without neutral class	2	0.53
Rough classification with neutral class	3	0.46
Detailed classification without neutral class	4	0.40
Detailed classification with neutral class	5	0.34

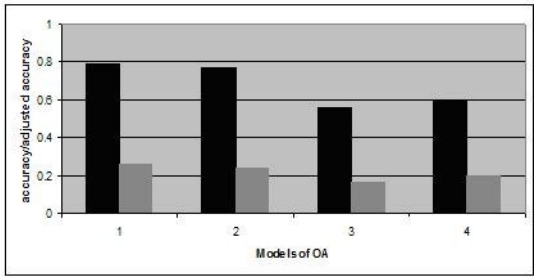


Figure 3:
Graphical illustration for Table 2

It is easy to see that in all cases the accuracy and adjusted accuracy of the rough opinion classification exceeds these values for the detailed opinion classification. In the case of the rough opinion classification the granularity of sentiments has no essential effect.

Evaluation of Opinion Analysis with the neutral category

In this series of experiments we studied the accuracy of decision-making on a set of all 50 documents. Table 13 contains the results of the calculation, also presented in graphic form in Figure 4.

The results show that the neutral category essentially decreases both accuracy and adjusted accuracy in comparison with the cases when the neutral category is absent. One of the principal reasons for such a situation is the relative small number of papers belonging to this category.

Evaluation of different rules for decision making on regression

In our previous experiments we used the rules for decision-making on regression presented in Tables 7–10. In this series of experiments we study OA with other rules.

Table 12: Values of accuracy and adjusted accuracy (the neutral category is absent)

Models of OA	Sentiment classification	Opinion classification	accuracy	Baseline	adjusted accuracy
1	rough (2 categories)	rough (2 categories)	0.79	0.53	0.26
2	detailed (4 categories)	rough (2 categories)	0.77	0.53	0.24
3	rough (2 categories)	detailed (4 categories)	0.56	0.40	0.16
4	detailed (4 categories)	detailed (4 categories)	0.60	0.40	0.20

Table 13: Values of accuracy and adjusted accuracy (the neutral category is present)

Models of OA	Sentiment classification	Opinion classification	accuracy	Baseline	adjusted accuracy
1	rough (2 categories)	rough (3 categories)	0.54	0.46	0.08
2	detailed (4 categories)	rough (3 categories)	0.52	0.46	0.06
3	rough (2 categories)	detailed (5 categories)	0.30	0.34	-0.04
4	detailed (4 categories)	detailed (5 categories)	0.42	0.34	0.08

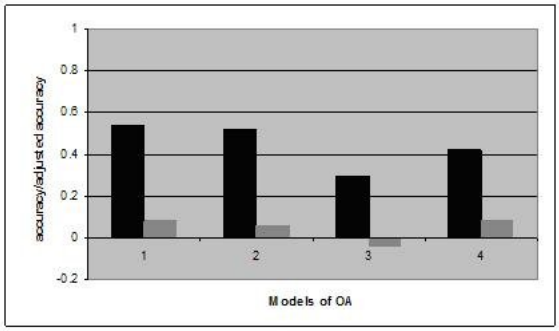


Figure 4:
Graphical illustration for Table 13

In the experiments we use all 50 documents, i.e. we consider the neutral category.

Tables 14 and 15 describes the rules with preference to the neutral category. Table 16 contains the results of the calculations and shows that there is no sense in preferences to the neutral category. It means that the linguistic variable for the neutral category is concentrated in a narrow interval near zero. Tables 17 and 18 describe the rules with preference to the polar categories adjacent to the neutral category. Table 19 contains the results of calculations.

The results show that the rules with preference to the polar categories and the detailed sentiment classification allow to obtain the best adjusted accuracy when

Table 14: Rules for decision-making with 3 categories of opinion (preference to the neutral category)

Regression value	Opinion
$R < -0.75$	Negative
$-0.75 \leq R \leq 0.75$	Neutral
$R \geq 0.75$	Positive

Table 15: Rules for decision-making with 5 categories of opinion (preference to the neutral category)

Regression value	Opinion
$R < -0.75$	Very negative
$-0.75 \leq R < -0.5$	Negative
$-0.5 \leq R \leq 0.5$	Neutral
$0.5 < R \leq 0.75$	Positive
$R > 0.75$	Very positive

Table 16: Values of accuracy and adjusted accuracy (preference to the neutral category)

Models of OA	Sentiment classification	Opinion classification	accuracy	Baseline	adjusted accuracy
1	rough (2 categories)	rough (3 categories)	0.32	0.46	-0.14
2	detailed (4 categories)	rough (3 categories)	0.34	0.46	-0.08
3	rough (2 categories)	detailed (5 categories)	0.28	0.34	-0.06
4	detailed (4 categories)	detailed (5 categories)	0.3	0.34	-0.04

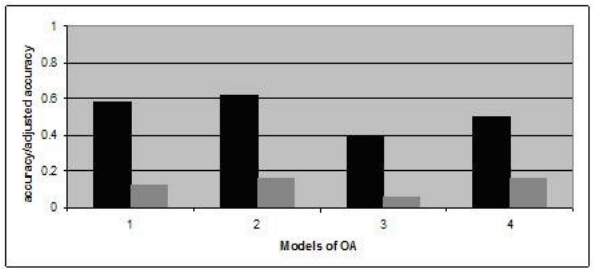


Figure 5:
Graphical illustration for Table 19

we deal with the neutral category. It concerns both the rough opinion classification and the detailed opinion classification.

Conclusion

Discussion

We completed OA of publications related to a given specific domain using very limited sentiment vocabularies. It was shown that in case of binary opinion classification regardless of the sentiment granularity the accuracy proved to be close

Table 17: Rules for decision-making with 3 categories of opinions (preference to the polar categories)

Regression value	Opinion
$R < -0.25$	Negative
$-0.25 \leq R \leq 0.25$	Neutral
$R \geq 0.25$	Positive

Table 18: Rules for decision-making with 5 categories of opinions (preference to the polar categories)

Regression value	Opinion
$R < -0.85$	Very negative
$-0.85 \leq R < -0.15$	Negative
$-0.15 \leq R \leq 0.15$	Neutral
$0.15 < R \leq 0.85$	Positive
$R > 0.85$	Very positive

Table 19: Values of accuracy and adjusted accuracy (preference to the polar categories)

Models of OA	Sentiment classification	Opinion classification	accuracy	Baseline	adjusted accuracy
1	rough (2 categories)	rough (3 categories)	0.58	0.46	0.12
2	detailed (4 categories)	rough (3 categories)	0.62	0.46	0.16
3	rough (2 categories)	detailed (5 categories)	0.40	0.34	0.06
4	detailed (4 categories)	detailed (5 categories)	0.50	0.34	0.16

to that obtained with large sentiment vocabularies and a very detailed sentiment classification -~0.8 (Pang and Lee, 2002; Brooke et al., 2009).

We studied the sensibility of OA to the sentiment classification and to the opinion classification. We showed that the adjusted rules of decision-making on regression equation allow to obtain satisfactory results when we deal with the neutral category.

Future work

In the framework of existing model of decision-making we suppose:

- To test the sensibility of OA to size of sentiment vocabularies;
- To consider publications on the same topic (economic crisis) written in Russian and in English;
- To consider publications on other topics (culture, politics) in Spanish.

We suppose to study OA:

- with mixed categories, such as neutral-positive and neutral-negative;
- with so-called 'undefined' category when it is better to say 'I do not know' than to give a certain answer.

In the paper we used the simplest lineal regression model with respect to linguistic variable. We intend to construct more complex models using the technique of Inductive Modelling (Alexandrov et al, 2009).

References

Alexandrov, M., Blanco, X., Catena, A., Ponomareva, N. (2009): Inductive Modeling in Subjectivity/Sentiment Analysis (case study: dialog processing). In: *Proc. of 3rd Intern. Workshop on Inductive Modeling (IWIM-09)*. Poland, pp. 40–43.

Brooke, J., Tofiloski, M., Taboada, M. (2009): Cross-Linguistic Sentiment Analysis: From English to Spanish. In: *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*. Borovets (Bulgaria), pp. 50–54.

Pang, B.. Lee, L. (2004): A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL)*. Barcelona (Spain), pp. 271–278.

Pang, B., Lee, L. (2008): Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1–135.

Pang, B., Lee, L., Vaithyanathan, S. (2002): Thumbs up? Sentiment classification using Machine Learning techniques. In: *Proceedings of Conference on Empirical Methods in NLP.* , pp. 79–86.

Pinto, D. (2008): *On Clustering and Evaluation of Narrow Domain Short-Text Corpora*. (PhD thesis.) Valencia (Spain) : Polytechnic Univ. of Valencia.

Taboada, M., Anthony, C., Voll K. (2006): Creating semantic orientation dictionaries. In: *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*. Genoa (Italy), pp. 427–432.

Weka [on-line]. (2009): [cit. 2010]. Available at: `http://prdownloads.sourceforge.net/weka/weka-3-7-0jre.exe`.

Do Central Players Perform Better?

Eva Maria Eckenhofer
Centre of Applied Economic Research, Faculty of Management and Economics, Tomas Bata University Zlin, Mostní 5139, 760 01 Zlín, Czech Republic
e-mail: Eva_Eckenhofer@hotmail.com

Abstract

Social Capital, the outcome for individuals from networks with shared norms and values, has already been discussed as a driver for innovation and performance improvement. Social Capital is a resource embedded in social structures, which can be accessed as well as mobilised in purposeful actions. The functions of Social Capital are transparency, which reduces transaction costs by improving information flow, and rationalisation, which reduces uncertainty and increases flexibility leading to enhanced performance and innovation. There exist various theories about social resources and structures leading to Social Capital, discussing whether network closure or the absence of ties is the key to success. Nevertheless little is known about the relation between network centrality and Social Capital. Therefore this paper aims to contribute to the discussion by analysing in a case study the structural position of actors who are rich in Social Capital. Additionally it will be assessed if those actors who are central in the social network are the ones with the highest performance. This study was based on a survey of 170 students from a Czech University who form three different networks. For the detection of Social Capital a procedure developed and tested in the European Values Study Surveys was applied and the relational data has been analysed by social network analysis using UCINET.

Key words social networks; social capital; network structure; performance

Biographical note

Mag.(FH) Eva Maria Eckenhofer, graduated in Media Management from University of Applied Sciences in St.Pölten (Austria) in 2008 and works now as a PhD Student on Tomas Bata University in Zlín on the Faculty of Management and Economics on her dissertation in the fields Social Capital, trust and organisational networks. The title of her dissertation will be: 'Strategic Networking as a Management-tool'.

Social Capital and its measurement

Social Capital has become a scientific buzzword and the discussions about its definitions, forms and attempts to measurement are widespread. Nevertheless it is questioned whether Social Capital is actually a form of capital (Halpern, 2005). What is known for sure is that the importance of this form of capital is imbedded

in social networks and that its importance increases under imperfect competition. (Burt, 1992)

While Lin defined it as a 'resource embedded in a social structure that are accessed and/or mobilised in purposive actions', (Lin et al., 2008) Putnam sees that 'the central idea of social capital is that networks and associated norms of reciprocity have value'. (Robert D. Putnam, 1996). Another component in building this resource is trust, which is defined as an expectation that arises within a community of regular, honest and cooperative behavior based on commonly shared norms (Fukuyama, 1995). Tsai divided Social Capital into three dimensions, the structural, where the contacts of an actor are located, relational, where the assets such as trust and trustworthiness are rooted, and a cognitive dimension which includes a shared code and vision (Tsai, 1998). Therefore it can be summarised that in general Social Capital is a resource which is embedded in social networks based on trust and specific norms.

How can these resources be attained or even measured? We know that due to participation in associations individuals are likely to change their values and preferences (Paxton, 2002), and trust and civic-minded behavior emerge by involvement in formal and informal groups and associations (Putnam, 1996). This can be explained by self-enforcing agreements which are reached in repeated interactions and lead to trust within the group, but also to civic behavior in general (Knack and Keefer, 2003).

In the UK this principle was taken in order to grow community involvement by 'corporate and employee volunteering'. The benefits are not only leading towards a trusting and networking community, but moreover it exhibits benefits for every individual as well as the companies (Muthuri et al., 2009). These benefits are intangible (reputation, knowledge) as well as tangible (financial and material). Moreover a shared vision helps an organisation to develop Social Capital and combine resources (Tsai, 1998).

Social capital leads to benefits on multiple levels, on an individual, group and community level (Paxton, 2002), but in general it contains structural and action-oriented elements (Lin et al., 2008, p. 58) and the returns can be categorised into returns to instrumental action and returns to expressive action (Lin, 1999). Returns on instrumental actions are economic, political and social return. Economic return can be the increase of turnover due to a new customer. Political return is, e.g., the influence on a legislative change and social return can be a contribution to a better reputation. Return on expressive action enforces and secures one's resources against possible losses. Moreover these effects make a positive contribution to one's physical and mental health as well as life satisfaction (Halpern, 2005), which goes with Cooke's statement: 'Human Capital is judged by individual income, while social capital is judged by quality of life.' (Cooke, 1999). Following Cooke (1999), the benefits lead back to embeddedness (communication benefits, integration and synergy) as well as to autonomy (integrity, linkage).

This leads to the assumption that actors with higher Social Capital have the possibility to perform better than other actors with lower Social Capital as they can mobilise higher amounts of resources which lead to returns on instrumental or expressive action.

Attempts at measuring this kind of capital, lead from Lin's Position-Generator,

where professions in one's ego-network are queried (Lin et al., 2008, p. 77), to Snijder's Resource-Generator, where specific services in one's Ego-Network are asked for (Van Der Gaag and Snijders, January 2005). Van der Gaag and Snijders argue this as follows: 'Here, we concentrate on measuring social capital within the 'access' perspective, and define social capital as the collection of all potentially available network members' resources.' (Van Der Gaag, Snijders January, 2005).

Another approach comes from Beugelsdijk and Van Schaik who combine general and institutional trust, group-membership, volunteering, free-time behavior and trustworthiness, in order to generate a Social Capital Index, by taking data from the European Value Studies (Beugelsdijk, 2005; Beugelsdijk and Van Schaik, 2002, 2005; Beugelsdijk et al., 2004). This fits to the idea that 'Social Capital is a communal property involving civic engagement, associational membership, high trust, reliability and reciprocity in social networks.' (Cooke, 1999) Moreover, as Social Capital has been defined as a resource embedded in a social structure (Lin et al., 2008, p. 58), it measures the investments made into one's social network in general. These investments are done over a longer time, as social networks and trust needs time to be built and tested (Fukuyama, 1995), so that when needed they are a channel for information and resource flow and therefore an entrance ticket for future options (Tsai, 1998; Lechner, 2003).

Network Positions and their effects

From numerous studies it is know that there exists a connection between Social Capital and economic performance (Beugelsdijk, 2005), between Social Capital and the quality of governance and economic growth (van Bouma, 2005), and between education and Social Capital as well as between Social Capital and health (Halpern, 2005).

Also concerning the influence of the structure of a social network and the positions of actors within it we know that productivity (Granovetter, 2005), resources-access (Lin et al., 2008, p. 76) , knowledge-transmission (Halpern, 2005) and innovation (Cooke, 1999) are influenced. Burt classifies these benefits into information and control benefits and ascribes the advantages of actors in a social network to their position as brokers, next to structural holes (Burt, 1992). Coleman sees the reason network benefits in the network closure (Coleman, 1988) and Granovetter ascribes benefits to the type of the actors ties' (Granovetter, 2005).

Within a network, specific structural positions can be identified which all have different characteristics and opportunities due to their location in the network. Central connector, boundary spanner, information broker and peripherical specialist (Cross, Prusak, 2002), or broker, consultant, gatekeeper, representative and liaison (Hanneman, 2007), as they can be analysed in the Social Network Analysis Software UciNet, can be distinguished. Following Cross, central connectors link most people in a network, boundary spanners link different network parts, information brokers are local stars in a network and peripherical specialists are consulted for specialised information (Cross, Prusak, 2002). Due to their structural position these actors provide certain benefits for themselves, which leads to the idea that they are able to perform due to their position in some way better. Due to their structural characteristics it is possible to find them within a social network,

though we do not know anything about their general characteristics. This leads to a list of questions which shall be discussed in this paper using data from a case study.

Actors which are central in a network, central connectors, are, because of number and type of their contacts, more central within the network and therefore it can be assumed that they have the possibility to get access to a broader field of information. This could provide them with an advantage leading to better performance, compared to those actors who are not so centrally positioned. Therefore it will be asked in the scope of this paper whether there is a connection between centrality within a network and the performance of the actor having a central position.

Another interesting question is whether those actors having a central position within a network are also those who are more likely to have higher resources in Social Capital. It has been discussed above that trust, civic engagement and trustworthiness are main components of Social Capital. It can be assumed that an actor who trusts more is trustworthy and more engaged in society in general and is also more likely to be social and connecting within a specific network.

Trust and its influence

Trust can be defined on a general network or societal level as 'Expectation that arises within a community of regular, honest and cooperative behavior based on commonly shared norms on the part of other members of that society', (Fukuyama, 1995) but also at an individual relationship level as an attribute of a relationship, which is an expectation that alleviates the fear that the other one could behave opportunistically. Trustworthiness on the other side is an attribute of an individual (Tsai, 1998).

There exist different kinds of trust, the basic, simple one as in a friendship, the blind trust to a superior and the authentic trust based on skills and relationship (Dervitsiotis September, 2006). Trust is built over time, through interaction and evaluation on integrity (ethical attitude), benevolence (goodwill) and competence (ability) (Becerra, Huemer, 2002). The basis of building trust is interpersonal communication and proximity in psychological, cultural, social and physical dimensions (Becerra, Huemer 2002), (Lechner, 2003), (Gössling, 2007, 12:5). As proximity is a criterion of trust it can be assumed that high trust is going along with high proximity within a network, or on the opposite a low trust level goes together with lack of proximity and therefore a low network density.

The effects of trust on the networks in which it arises as well as on the actors within trusting networks have been studied. Trust is said to enable more efficient operating processes (Dervitsiotis September, 2006), matters in the effectiveness of exchange relations, especially in inter-organisational relationships (Becerra, Huemer, 2002).

On a societal level higher trust increases investment and growth (Van Schaik, 2002), and on the relationship level trust is associated with greater open communication, lower emotional conflict, faster decision-making and greater willingness to take risks. As trust reduces the complexity, the need for constant surveillance and the constraint of opportunism, it leads to a decrease of transaction costs for individuals as well as for companies (Becerra and Huemer, 2002). This is possi-

ble as trust reduces monitoring costs and enables heuristic-based decision-making (Uzzi, 2008). Another positive influence of trust is that information exchanges are more proprietary and tacit, and that it reduces therefore the information asymmetry between parties. As trustful relations within a network are said to increase information flow and lower monitoring cost, it can be assumed that within a network of a higher trust level, also the performance will be better.

Assumptions and Methodology

In Part 1 to 3, several assumptions, based on scientific literature and studies which have been done, shall be discussed and enlightened by a small survey which was conducted at Tomas Bata University between December, 2009 and May 2010:

Assumption 1: Higher Social Capital of an actor is connected to higher performance.
Assumption 2: Central position of an actor is connected to higher performance.
Assumption 3: A higher level of Social Capital is connected to higher centrality.
Assumption 4: The level of trust of a network is connected to its overall performance.
Assumption 5: A higher level of trust of an actor is connected to higher performance.

For analysing these assumptions three groups of students at Tomas Bata University Zlín (CZ) have been asked to fill in a questionnaire. A total of 170 students filled in the questionnaire, from this number 41 were second-year students who subscribed to a Desktop-Publishing lecture, 56 first-year students who followed statistics lecture and 73 were PhD students of the faculty of Management and Economics on Tomas Bata University.

The questionnaire contained questions about students' relationships to their colleagues, the number of languages they spoke, if they had already been abroad for more than three months, and questions linked to social capital which were used in the European Value Studies (Beugelsdijk, Van Schaik). First it was asked: 'Generally speaking, would you say that most people can be trusted or that you can't be too careful in dealing with people?' Then the students were asked to evaluate their level of trust in institutions like a church, parliament, press etc. Then they were asked if they were a member or volunteer in certain organisations, how they spend their free-time and finally, in order to evaluate their trustworthiness, they were asked to estimate whether certain statements like 'Claiming state benefits which you are not entitled to' are always, sometimes, rarely or never justified.

From the relational questions social network analysis of the whole network using Ucinet has been done. Social network analysis is a social ethnological method which can be used to measure and visualise the social structure of a group as a whole and the social embedment of its individuals / actors (Schnegg and Krenn, 2009; Jansen, 2006; Wasserman and Faust, 2008). The focus of a social network analysis can be a single actor or an aggregate of persons – whole groups, as it has been done in this study. The components are the social relations between these actors, which can be based on kinship or friendship. In our case, communication, going-out, advice-seeking and lending-money relations have been collected.

Using the aggregated total network, centrality and prestige measures of the whole network and from individual actors has been calculated. These concepts are based on the idea that the actor who has many ties within the network is more central and therefore more visible. Prestige measures show actors who can influence the network. It is a contribution to social capital, as more prestigious actors have more access to resources. As there is not a single measure which describes centrality best, all three major centrality measures, degree-based, closeness-based and betweeness-based centrality have been calculated in order to correlate them later with the performance data of the students (Hanneman, 2007).

Degree-based centrality is measured by the outdegree of an actor, which computes all outgoing relations to other actors in the case of an asymmetric and directional network. For a symmetric and nondirectional network all relations are computed. Closeness-based centrality measures not only the direct but, moreover, the indirect relations to other actors (path distances). The closeness of an actor is measured by the reciprocal of the sum of all path distances of an actor. Betweeness-based centrality has a different logic as degree-based and closeness-based centrality as it starts from a dyad and computes the shortest path distance from one to another, called geodesic. The idea behind it is the probability that communication from actor *a* to actor *b* will run over actor *c*. The ratio of the number of geodesics between *a* and *b* going through *c* to the total number of shortest paths between *a* and *b* is computed in order to get the betweeness-based centrality (Wasserman and Faust, 2008; Jansen, 2006). The next step was to run factor analysis, which can be used to reduce the number of variables, to detect structure in the relationships between variables and to classify them. (StatSoft, 2010). Therefore, as factor analysis can be applied for data reduction, the five trust, trustworthy and public involvement questions have been reduced to one variable called 'Social Capital' as it has been done by before Beugelsdijk with three variables in order to create a social capital index using data from the European Value Studies (Beugelsdijk and Van Schaik, 2005; Van Schaik, 2002). After this step correlations were done from the new variable 'social capital' as well as the original ones, the performance variable, the number of languages, being abroad and the centrality measures of the actors.

Moreover the relationships between the average performance, level of Social Capital and trust, centralisation and density of the whole networks were computed. Finally the structural position of the better performing actors in average of their grades as well as Social Capital, have been analysed qualitatively.

Analysis

Statistical Analysis

Correlations

After aggregating the different relations in every network, the centrality measures degree, closeness, reach and betweeness have been calculated in Ucinet for every actor from each network. For analysis we used the values number of languages, being abroad, average degree, trust, institutional trust, involvement, free time and trustworthiness to correlate them with the centrality measures, calculated in Ucinet.

For correlating other values with the performance measure average grade we had to exclude the network of PhD students as they do not get any grades and their performance could have been measured only by the number of their publications.

On a significant level of p-value under 0.05 we found several weak correlations and one moderate correlation. The moderate correlation we found between involvement and free time with r=0.565.

On a weak level being abroad correlates with the number of languages an actor speaks. The number of languages someone speaks correlates with the variable inCloseness. Another interesting weak negative correlation has been found between the average grade and being abroad. The negative direction can be explained as a lower average grade means a better performance than a higher average grades. The grades are measured on a scale ranging from 1 to 3. Also weak and negatively correlates the average grade with the centrality measures degree, share and reach. Surprisingly also negatively weak is the correlation between institutional trust and inCloseness, but positive with outCloseness. Another negative weak correlation has been found between the value trustworthiness and outCloseness and outward Reach. A weak positive correlation again has been found between the values for free time behaviour and the centrality measure degree. A significant correlation has neither been found between the five Social Capital Values, nor between trust and performance.

Also on a meta-level comparing the average of the three networks there was no correlation found between density, centralisation and clustering and social capital. One strong correlation from 0.997 at 0.05 p-value has been found between the average level of trust and social capital, which is logical as Social capital is based on trust.

Factor Analysis

Before doing Factor Analysis from all five Social Capital values as Van Schaik (2002) proposed, we ran Factor Analysis from Trust, active and passive Membership as Beugelsdijk and Van Schaik did in 2005 (Beugelsdijk, 2005).

Our factor loadings from these three variables are 0.246 for trust, 0.654 for passive Membership and 0.827 for active membership. The result of their factor loadings were 0.49 for trust, 0.75 for passive membership and 0.89 for active group membership.

Our factor loadings were about 0.2 points smaller which can be explained by the difference in n, while we were calculating from 170 items, Beugelsdijk and Van Schaik used the database from the European Value Studies and had supposedly many more items. Nevertheless the rank and the differences in factor loadings were similar.

Even as the correlations of the five social capital values were not significant we proceeded to the next step to do factor analysis of these values. By calculating one factor we got a lower p-value as for calculating two factors, which supposes on the one hand that two different factors would be a more adequate explanation. On the other hand, some of the factor loadings are higher than 0.4, which is quite good and, moreover, in total these five variables describe 66.5% of all variance, while p-value suggests that the null hypothesis is correct. Therefore we decided to

Table 1: Correlations of the five Social Capital values.

	Trust	Instit. Trust	Involvement	FreeTime	Trustworth.
Trust	1.00000000	0.14766883	0.15876757	0.11571389	0.02182192
Instit.Trust	0.14766883	1.00000000	0.10386980	-0.04641903	0.05485412
Involvement	0.15876757	0.10386980	1.00000000	0.21750242	0.06182872
FreeTime	0.11571389	-0.04641903	0.21750242	1.00000000	-0.05979438
Trustworth.	0.02182192	05485412	0.06182872	-0.05979438	1.00000000

Source: own

proceed with one factor called Social Capital.

Table 2: Factor Loadings for Factor Analysis the five Social Capital values, calculating one factor

Loadings:	Factor1	Loadings:	Factor1	Factor2	
Trust	0.351	Trust	0.135	0.288	
Instit. Trust	0.357	Instit. Trust	0.994		
Involvement	0.430	Involvement		0.501	Source: own
FreeTime	0.407	FreeTime		0.432	
Trustworth.	0.252	Trustworth.	0.176	0.140	
The p-value	0.624	The p-value	0.656		

The two different factors provided by factor analysis derive from different variables. Factor 1 derives mainly trust Institutional Trust followed by Trust and Trustworthiness. This factor could be described as an overall Trust Value. The second one is mainly based on Involvement (Group Membership) and Free Time Behaviour, complemented by trust and trustworthiness. This factor can be described as a societal value of an actor.

In order to analyse if Social Capital has an influence on the centrality or performance of an actor, we calculated a Social Capital Value based on the factor loadings. Contrary to Beugelsdijk and Van Schaik, we did not rescale it, as the primary use was to calculate the size of actor nodes attributes based on Social Capital Value, similar as it has been done with performance.

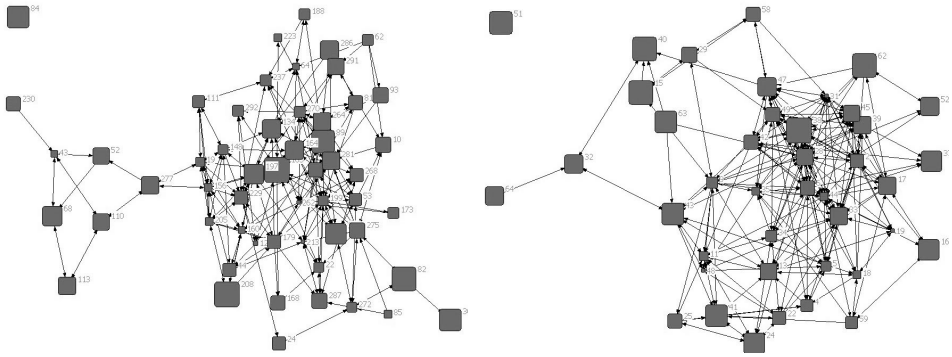
Correlating the new Social Capital Value with the centrality measures gave a low but on 0.05 p-value significant correlation of 0.225 with InCloseness. Also Degree and Reach-centrality were on a similar level significant.

No Significant correlation was found between Social Capital Value and the performance of a student.

Qualitative Analysis

For qualitative analysis the aggregated social networks of the statistic-students and desktop-publishing students have been displayed by Netdraw. The relational ties have been kept bidirectional and the graph-theoretical layout of the network was generated by spring embedding, an algorithm that uses iterative fitting to locate the points to each other according to their smallest geodesic distance.

Figure 1: Aggregated Social Network of Statistics (left) and Dtp-Students(right), node-size according to average-grade



Source: own

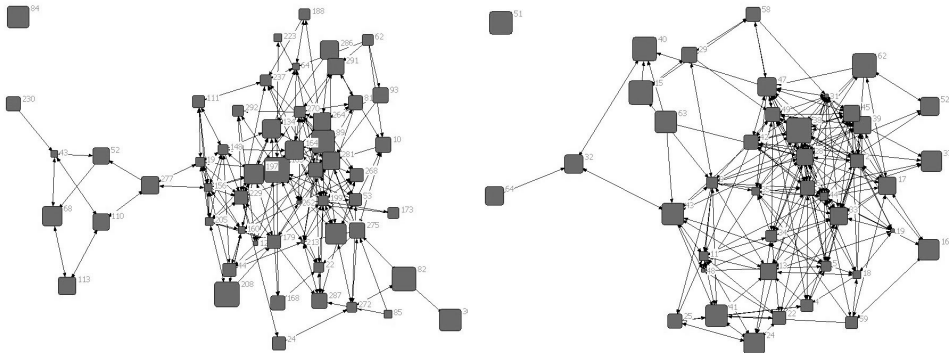
In Figure 1 the node size in the students' networks has been calculated according to their average grade. As the average grade is measured on a scale from 1 to 3, a better student is indicated by a smaller node within the network.

Comparing these two networks it is visible that in both networks we can find smaller nodes, which indicate that those are good students, in the centre and having many ties to others. This observation is a visible confirmation of the correlation done prior. In the network of statistic students on the left side we see that the actor with the number 43 is really small and has three reciprocal ties within this sub network. In the right part of the statistics network we can find several small (above average) students who are connected to many other nodes, such as 19, 156, 205, 262, 64 and many more.

On the right side in the network of desktop publishing students we can find five really small nodes which are densely connected to others. An interesting example is actor 11 and 48, who are both small, highly connected and close to each other.

In those two networks the node size has been calculated based on the social capital level of the student. Therefore a bigger node indicates a higher level of social capital, which has been calculated based on the factor loadings from the factor analysis. Not as obvious as for performance, but still we can find bigger nodes in the centre. Several 'big' nodes, which are rich in social capital we can find next to global payers in the role of an insider or hub. Actor 277 of the desktop publishing network is building the link between the sub network and the main network and has a considerable bigger larger node size, and therefore higher level of social capital, than the actors around him. In the network of Statistic-Students actor 9 is a similar example, he is connecting the main network and the 'outsiders' of the network. Also actor 29, who is connecting the outsiders of the network, has many ties and is bigger than the actors around him. An exception from this is actor 13, who is, while being small, densely connected within the main network.

Figure 2: Aggregated Social Network of Statistics (left) and Dtp-Students(right), Node-size according to Social Capital



Source: own

In Figure 3 the size of the node has been calculated considering both the levels of social capital and performances. This does not show much difference; just a small tendency is visible that bigger ties are more central, this is more the case in the network of desktop publishing students than in the network of the statistic students, where we can find a very small node totally in the centre and several big nodes located at the border of the network. At the network of desktop publishing students, we have a smaller sub-group at the left side and many bigger ones in the right part of the network, though still in the centre of the network there are three small nodes.

Discussion

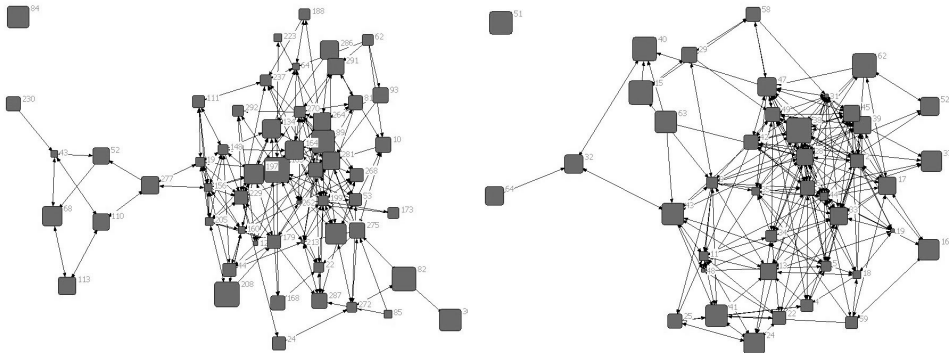
After statistical and qualitative analysis of the data collected in a small survey, we can now discuss the findings under consideration of the assumption done beforehand in the literature review.

The first assumption made was that higher Social Capital of an actor is connected to higher performance. In the student networks we analysed, we could not find any correlation between Social Capital and performance. It can be taken into consideration, whether the kind of Social Capital measurement or the measuring of performance might be the reason, or if these two variables do not affect each other.

The next assumption was that the central position of an actor is connected to a higher performance. Analysing the networks of statistic and desktop publishing students we found a significant correlation indicating that centrality is to some extent connected to performance.

Also for assumption three, concerning the connection between social capital and centrality, we found proof in our data. A weak, but significant correlation is between social capital and centrality, which leads to the conclusion that in fact in

Figure 3: Aggregated Social Network of Statistics (left) and Dtp-Students(right), Node-size according to Social Capital and average grade



Source: own

the networks analysed central players perform better by means of average grade and social capital.

Proof of connection between trust and performance was neither found on an individual, nor on an aggregate level. Therefore assumptions four and five were rejected by our data.

Interesting findings were the correlations between being abroad and the number of languages, as well as the number of languages and closeness and the average grade and being abroad. These correlations suggest the conclusion that learning languages and spending some time abroad in a foreign country has a positive influence on the average grade and being in a central position within a network.

The measurement of social capital was another important part of our survey. From the literature review the methodology proposed by Beugelsdijk and Van Schaik was integrated into our study. Adopting the questions used in the European Value Studies, where the findings are used to compare countries, for analysing Social Capital on an individual level, an experiment was done first in the field of Social Network Analysis and Social Capital Studies. Van Schaik proposed the four dimensions of Social Capital: Interpersonal trust, Institutional trust, Participation in civic society (formal and informal) and trustworthiness (Van Schaik, 2002). Nevertheless as the use of these dimensions for generating a Social Capital Index was not found before in scientific literature, factor analysis from trust, active and passive membership as Beugelsdijk and Van Schaik did in 2005 (Beugelsdijk, 2005), was done beforehand, in order to see if the factor loadings are comparable even in a smaller amount of data. As the factor loadings were comparable, we ran Factor Analysis from all five dimensions, five values. The problem was that the only significant but, therefore, moderate correlation was between involvement and free time with r=0.565. This finding addresses the questions about formal or informal membership and the question about free time behaviour, are looking into the same dimension: Participation in civic society.

Another insight was that trustworthiness does not significantly contribute to Social Capital, though the question is whether this is really the case, or whether this effect results from social expectancy and cheating. Nevertheless a weak connection between the variable called Social Capital, calculated from the factor loadings from factor analysis, and centrality within the network could have been found.

Conclusion

Disregarding the uniqueness and the novelty of the findings in this survey, this survey has some limitations due to the size. At the centre of the analysis were three different networks of students, with a total number of 170 students, which does not allow any generalisation. Another limitation is the small size of the correlations found in the data, which is on the one hand clearly disputable, but on the other hand obvious as the performance of a human being, especially the examination performance of a student will never depend on one single variable.

It is a matter of further studies to confirm the findings from these three networks on a broader scope and evaluate if the connections are comparable to other student groups, student groups in different countries, or even to other types of networks.

The connection between Social Capital and centrality within a network, as well as between centrality of an actor and his performance, should be analysed in detail in the future, as the social-capital-questions used in this study could potentially be posed in job. The social capital dimensions and the questions used for collecting them, which were adopted for this survey from the European Value Studies, have to be tested again and analysed to see if a social capital index generated from its values is justified and comparable.

In business also the centrality of employees within the intra-organisational network can become an evaluation criterion, if a significant connection between centrality and performance could be proved in general. Studying this is always connected with the problem of defining performance. As we had difficulty in this survey to define the performance of a PhD student, also in business the performance of an employee cannot always be measured by a single variable.

Another consideration for further research is to determine whether the number of languages a person speaks or spends some time abroad also influences other fields.

In summary we can say that in the survey conducted in the scope of this paper central players do perform better, by means of average grade and Social Capital.

References

Becerra, M.; Huemer, L. (2002): Moral character and relationship effectiveness: an empirical investigation of trust within organizations. In: Heidi von Weltzien Hoivik (Hg.) *Moral leadership in action: building and sustaining moral competence in European organisations.* Glos : Edward Elfar Publishing, pp. *.

Beugelsdijk, S. (2005): Differences in Social Capital between 54 Western European Regions. *Regional studies, S.*, Vol. 39, No. 8 (2005), pp. 1053–1064.

Beugelsdijk, Sjoerd; Groot, Henri L. F. de; van Schaik, Anton B. T. M. (2004): Trust and economic growth. A robustness analysis. *Oxford economic papers*, Jg. 56, H. 1, pp. 118–134.

Beugelsdijk, Sjoerd; Van Schaik, Ton (2001): *Social capital and regional economic growth* [on-line] Center for Economic Research. [cit. 2010] Available at: http://greywww.kub.nl:2080/greyfiles/center/2001/102.html.

Beugelsdijk, Sjoerd; Van Schaik, Ton (2005): Social capital and growth in European regions. An empirical test. *European Journal of Political Economy*, Jg. 21, H. 2, pp. 301–324.

Burt, Ronald S. (1992): *Structural holes. The social structure of competition* Cambridge (Mass) : Harvard Univ. Press.

Coleman, J. S. (1988): Social Capital in the Creation of Human Capital. *The American Journal of Sociology*, Jg. Vol. 94, H. Supplement: Organizations and Institutions: Sociological and Economic Approaches to the Analysis of Social Structure., pp. *.

Cooke, P. Wills D. (1999): Small Firms, Social Capital and the Enhancement of Business Performance Through Innovation Programmes. *Small Business Economics*, Jg. 13, pp. 219–234.

Cross, Rob; Prusak, L. (2002): The People who make organizations go-or Stop. In: *Harvard Business Review, H. June.*. , pp. *.

Dervitsiotis, K. (September 2006): Building Trust for Excellence in Performance and Adaptation to Change. *Total Quality Management*, Jg. Vol. 17, H. 7, pp. 95–810.

Fukuyama, Francis (1995): *Trust. The social virtues and the creation of prosperity.* New York : Free Press.

Gössling, T. (2007): Proximity, trust and morality in Networks. *European Planning Studies*, Jg. 2007, 12:5, pp. 675–689.

Granovetter, Mark (2005): The impact of social structure on economic outcomes. *The Journal of Economic Perspectives*, Jg. 19, H. 1, pp. 33–50.

Halpern, David (2005): *Social capital* Oxford : Polity Press.

Hanneman, Robert A. (2007): *Introduction to social networks.* .

Jansen, Dorothea (2006): *Einführung in die Netzwerkanalyse. Grundlagen, Methoden, Forschungsbeispiele [Lehrbuch]* 3. überarb. Aufl. ed. Wiesbaden : VS Verl. für Sozialwiss.

Knack, Stephen; Keefer, Philip (2003): Does social capital have an economic payoff? A cross-country investigation. In: *Democracy, governance, and growth.* , pp. 252–288.

Lechner, Ch.; Dowling M. (2003): Firm networks: External relationships as a source for the growth and competitiveness of entreprenеurial firms. *Entrepreneurship and Regional Development*, Jg. 15:1, pp. 1–26.

Lin, Nan (1999): Building a Network Theory of Social Capital. *Connections INSNA*, Jg. 22(1), pp. 28–51.

Lin, Nan; Cook, Karen S.; Burt, Ronald S. (2008): *Social capital. Theory and research* 4 ed. New Brunswick : Aldine Transaction.

Muthuri, J.; Matten, D.; Moon, J. (2009): Employee Volunteering and Social Capital: Contributions to Corporate Social Responsibility. *British Journal of Management*, Jg. Vol. 20, pp. 75–89.

Paxton, Pamela (2002): Social Capital and Democracy: An Interdependent Relationship. *American Sociological Review*, Jg. 67, H. 2, pp. 254–277.

Robert D. Putnam (1996): The Strange Disappearance of Civic America. *The American Prospect*, Jg. 24, H. Winter, pp. 34-38.

Schnegg, M.; Krenn, K. (September 2009): Network Analysis in the Humanities and the Social Sciences Herausgegeben von Summerschool Modul 1. (Theorie und Geschichte) Skript. Trier.

StatSoft, Inc. (2010): Electronic Statistics Textbook Available at: http://www.statsoft.com/textbook/.

Tsai, W. Ghoshal S. (1998): Social Capital and Value Creation: The Role of Intrafirm Networks. *The Academy of Management Journal*, Jg. Vol. 41, H. 4, Aug., pp. 464–474.

Uzzi, Brian (2008): Social structure and competition in interfirm networks. The paradox of embeddedness. In: *Small business and entrepreneurship.* , pp. 400–432.

van Bouma, J.; Soest D.; Bulte E. (2005): *Trust, trustworthiness and cooperation: Social capital and community resource management* .

Van Der Gaag, Martin; Snijders, Tom A. B. (January 2005): The Resource Generator: Social capital quantification with concrete items. *Social Networks*, Jg. 27, H. Issue 1, pp. 1–29.

Van Schaik, Ton (2002): Social Capital in the European Value Study Surveys. In: *Country paper prepared for the OECD-ONS International Conference on Social Capital Measurement.* London, pp. *.

Wasserman, Stanley; Faust, Katherine (2008): *Social network analysis. Methods and applications* 7th print Cambridge : Cambridge Univ. Press.

The preparation of this paper was financially supported by the Internal Grant Agency of Faculty of Management and Economics on Tomas Bata University, project No. IGA&62/FaME/10/D.

The Study of Sentiment Word Granularity for Opinion Analysis (A Comparison with Maite Taboada works)

Olga Kaurova
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: kaurovskiy@gmail.com

Mikhail Alexandrov
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: malexandrov@mail.ru

Natalia Ponomareva
Statistical Cybermetrics Research Group, School of Computing and IT, University of Wolverhampton, Stafford Str, WV1 1SB Wolverhampton, UK
e-mail: nata.ponomareva@wlv.ac.uk

Abstract

Sentiment Analysis (SA) is an area of NLP related to automatic evaluation of people's opinions and their attitudes to various objects and events. Nowadays OA has become an important part of Social Network Analysis, and researchers suggest different tools for solution of this problem. The semantic orientation calculator (SO-CAL) developed in Maite Taboada's group is one such effective tool, which uses dictionaries of sentiment words over a detailed sentiment scale (5 positive and 5 negative levels). In the paper we study the influence of granularity levels of sentiment words on the accuracy of sentiment classification in order to verify the possibility of using lesser granularity without a substantial decrease in performance. We exploit one- and two-parameter linear regression models as a classification method and product reviews of different categories (books, cars, movies, etc.) as a corpus. The results show that there is no significant difference between one-/ and two-parameter models; neither is there a need for a fine-grained granularity of sentiment.

Key words opinion analysis; sentiment classification

Biographical note

Olga Kaurova graduated from the Saint Petersburg State University as a specialist in theoretical and applied linguistics in 2009. She is currently a graduate student at the Autonomous University of Barcelona in Spain (International Master's Program in Natural Language Processing & Human Language Technology. Her area of research interests – Sentiment and Subjectivity Analysis, Language Acquisition.

Mikhail Alexandrov is a member of the fLexSem Research group at the Autonomous University of Barcelona in Spain. He is a professor of the Academy of National Economy under Russian Government. He is an applied mathematician and author of numerous publications related to mathematical modelling and natural language processing. His current topics of research are machine learning (inductive modelling, clustering) and internet technologies (social networking).

Natalia Ponomareva is a PhD student at the University of Wolverhampton. She received her master's degree from the Technical University of Valencia. Her research interests include Sentiment, Sentiment Transfer and Machine Learning for NLP. She is author of more than 10 scientific publications in international conferences and journals.

Introduction

We always have to deal with subjectivity in our everyday life, have to take into consideration other people's opinions, and now with the growth of the WWW we get a quick and easy access to a great quantity of subjective information – opinionated texts: users' reviews, forums, blogs, etc. Analysis of this opinionated web content is becoming increasingly important both for individual and for business aims: for example, consulting consumer reports when choosing a brand of washing machine to buy, or monitoring the company's efficiency and satisfaction of its customers. Many online shopping sites, e.g. Amazon and eBay, give customers the possibility to leave their comments and reviews of the products they purchased. Moreover, there are even special sites, devoted to user's opinions: epinions.com and others. Thus, easy access to subjective data, on the one hand, and their large quantities and low level of order, on the other hand, determine the rapid development and great importance of Sentiment Analysis, which nowadays occupies a significant place in Natural Language Processing.

Related work

Sentiment analysis is a broad area of NLP, which concerns the automatic determination of text subjectivity (whether a text is objective or subjective), polarity (positive or negative) and sentiment strength (strongly or weakly positive/negative). One of the main tasks of sentiment analysis is a binary sentiment classification which aims to assign to an opinionated document either an overall positive or an overall negative opinion (sentiment polarity classification or polarity classification). There are two different approaches to achieving this aim: a lexical (lexicon-based) approach (Turney, 2002) and a machine learning approach (Pang et al., 2002). The machine learning approach uses collections of labelled texts as training data in order to build automated classifiers. The lexical approach is based on semantic orientation (SO) lexicons (words with their semantic orientation) (Hatzivassiloglou and McKeown, 1997), and calculates overall sentiment by aggregating the values of those words presented in a text or a sentence. Besides polarity classification of documents, other sentiment classification tasks have been receiving lots of attention in the research community: sentiment classification of subjective expressions (Wilson et al., 2005; Kim and Hovy, 2004), subjective sentences (Pang and Lee, 2004) and topics (Yi et al., 2003; Nasukawa and Yi, 2003; Hiroshi et al., 2004). These tasks analyse sentiment at a fine-grained level and can be used to improve the effectiveness of sentiment classification, as shown in the study of Pang and Lee (2004).

Maite Taboada's SO-CAL, which the present study is based on, belongs to the lexical approach. It is an automated system which uses low-level semantic and syntactic information to calculate the overall polarity of texts. So-CAL uses SO-

Dictionaries, which are lists of manually-tagged sentiment words. The current version consists of four open-class dictionaries (nouns, adjectives, adverbs and verbs) and one closed class-dictionary of intensifiers. The integer SO value assigned to each word varies between -5 and 5. The calculation of the sentiment orientation of a text is accomplished, roughly speaking, by summing the values of all the words occurring, also taking into account negation, intensification and other language phenomena. The use of a 10-point scale (excluding zero) of SO seems to be a compromise between an attempt to capture clear differences in word meaning on the one hand, and the difficulty in assigning extremely fine-grained values to out-of-context words on the other hand. The numerical values were chosen to reflect both the prior polarity and strength of the word, averaged across likely interpretations (Brooke et al., 2009).

Problem settings

The present study is founded on the work of Maite Taboada's research group, namely on their Semantic Orientation Dictionaries and a corpus of reviews, which we were kindly provided with by them. Our main objectives are:

- To study the influence of the adopted sentiment granularity scale on the accuracy of sentiment classification within the framework of the regression model.
- To compare performances of one- and two-parameter models, where the former model takes into account a summed contribution of positive and negative words while the latter model considers positive and negative scores separately as independent variables.
- To analyse the accuracy of regression models for each of the object categories of the corpus.
- To construct an integral model (built on all categories) and test its applicability to individual categories.

Models for decision-making

Source data and vocabularies

The corpus that served as material for the present work was developed by Maite Taboada's research group (Taboada et al., 2006). It is a collection of 400 texts obtained from the website Epinions (`www.epinions.com`). The reviews are divided into 8 object categories: books, cars, computers, cookware, hotels, movies, music and phones. Each category contains a set of 50 reviews: 25 positive and 25 negative. The reviews vary greatly in length: from several phrases to several pages. All the texts are written in English. General characteristics of the corpus are given in the table below.

We use four manually-ranked SO Dictionaries (nouns, adjectives, adverbs and verbs), where the integer SO value assigned to each word varies between 5 and 5. The dictionary of intensifiers is left out for the fact that we do not take into account negation, intensification, modality, etc. in this study. In Table 2 we present the size of the dictionaries.

Table 1: Characteristics of the experimental corpus

Characteristics	Value
Total number of reviews	400
Number of categories	8
Number of reviews in a category	50
Review's sentiment value	+1 / −1

Table 2: The size of SO-CAL dictionaries

Dictionary	No. of Entries
Adjectives	2257
Adverbs	745
Nouns	1142
Verbs	903

Document parameterisation

To examine the influence of SO granularity we introduced 5 different models of roughening the granularity scale (Table 3). According to these models we modified SO values in SO Dictionaries and performed parameterisation of each review of the experimental corpus. Table 4 presents an example of an output file after applying our document parameterisation procedure implemented in Python (fragment from the category BOOKS).

Table 3: Models of sentiment contribution

Model	SO-scale	Modified SO-scale
Model 1	−5, −4, −3, −2, −1, 1, 2, 3, 4, 5	−5, −4, −3, −2, −1, 1, 2, 3, 4, 5
Model 2	[−5, −4], −3, [−2, −1], [1, 2], 3, [4, 5]	−3, −2, −1, 1, 2, 3
Model 3	[−5, −4, −3], [−2, −1], [1, 2], [3, 4, 5]	−2, −1, 1, 2
Model 4	[−5, −4], [−3, −2, −1], [1, 2, 3], [4, 5]	−2, −1, 1, 2
Model 5	[−5, −4, −3, −2, −1], [1, 2, 3, 4, 5]	−1, 1

Regression models

We chose linear regression as the principal method of our analysis for several reasons. First of all, when considering one-parameter regression model with joint contribution of positive and negative words, our method becomes similar to the approach implemented in SO CAL. Second, this model can easily be adapted to multi-scaled sentiment classification although in this work only binary classification was carried out. Finally, a small amount of training data does not allow the exploitation of more sophisticated machine learning algorithms.

One of the objectives of this study is to check whether a two-parameter regression model (or Separate model), where contributions of positive and negative words are considered separately as independent parameters, has any advantage

Table 4: Fragment of an output of the Python program for SO calculation

file	name	npw	nnw	PS 1	NS 1	PS 2	NS 2	PS 3	NS 3	PS 4	NS 4	PS 5	NS 5
No 01	txt	13	9	19	−14	15	−11	15	−11	13	−9	13	−9
No 02	txt	8	8	13	−22	10	−15	10	−13	8	−10	8	−8
No 03	txt	44	23	73	−44	56	−30	55	−28	45	−25	44	−23
No 04	txt	10	6	15	−8	10	−7	10	−7	10	−6	10	−6
No 05	txt	40	18	65	−27	48	−21	47	−21	41	−18	40	−18
...	...	...	...	...	...	...	...	...	...	...	...	...	...

Legend: npw = number of positive words; nnw = number of negative words;
NS = negative score; PS = positive score

over a one-parameter model (or Joint model), where contributions of sentiment words are summed up.

Formula (1) presents Joint and Separate models we are going to construct and compare.

$$\text{Joint Model: } F_j = A_0 + A_1(\text{PosScore+NegScore}) \tag{1}$$

$$\text{Separate Model} F_s = A_0 + A_1\text{PosScore} + A_2\text{NegScore},$$

where A_0, A_1, A_2 are unknown coefficients.

We build linear regression models for different levels of sentiment granularity in order to find out whether finer-grained sentiment scales can significantly improve the results of classification. Besides individual models for each object category (categorial models), an integral (or multicategorial) model based on the whole dataset is constructed. It is used to verify the possibility of applying the same model to all categories without a substantial decrease in performance.

Prior to model testing we apply some modification to the variables. First of all, in order to avoid the dependency of sentiment scores on text length we normalise them on the total number of sentiment-words in a review. In order to be able to compare models of different levels of granularity, we adjust their variables to the same scale, namely, (-1, 1), by introducing a scale factor. The resultant forms of variables for different models are presented in Table 5 (Np stands for the number of positive sentiment-words in the text, and Nn – that of negative words). Sentiment scores before and after application of normalisation and scaling are shown in Figures 1 and 2.

Table 5: Normalisation and scaling of coefficients for model testing

Model	Joint Sentiment Score	Separate Sentiment Scores	
1	(PS 1 + NS 1) / (N_p+N_n) / 5	PS 1 / (N_p+N_n) / 5	NS 1 / (N_p+N_n) / 5
2	(PS 2 + NS 2) / (N_p+N_n) / 3	PS 2 / (N_p+N_n) / 3	NS 2 / (N_p+N_n) / 3
3	(PS 3 + NS 3) / (N_p+N_n) / 2	PS 3 / (N_p+N_n) / 2	NS 3 / (N_p+N_n) / 2
4	(PS 4 + NS 4) / (N_p+N_n) / 2	PS 4 / (N_p+N_n) / 2	NS 4 / (N_p+N_n) / 2
5	(PS 5 + NS 5) / (N_p+N_n)	PS 5 / (N_p+N_n)	NS 5 / (N_p+N_n)

Legend: PS = positive score; NS = negative score

All constructed regression models are checked for their statistical significance by a global test (F-test) and tests on individual variables (t-test).

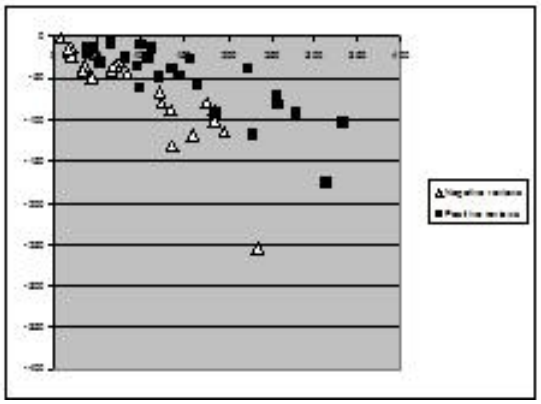


Figure 1:
Distribution of scores before normalisation

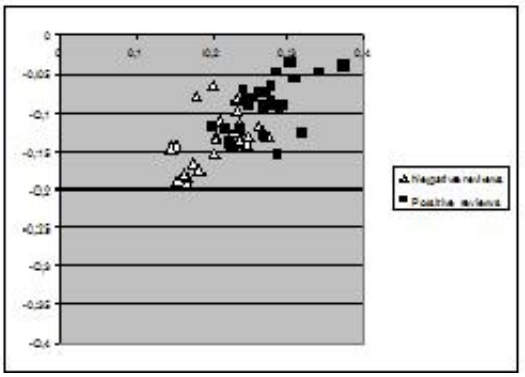


Figure 2:
Distribution of scores after normalisation

Model accuracy and model comparison

In as far as the scale of sentiment of reviews under consideration is binary: +/−1, the coefficients of determination R2 and the values of standard errors are not representative for the evaluation and comparison of regression models. Therefore we apply a cross-validation technique to estimate model accuracies (accuracy in our case means the probability of correct classification of a review as negative or positive). Taking into account the fact that the experimental data are not very large and the level of noise is high, the leave-one-out cross-validation method is used.

In order to compare model accuracies the z-test is used (it is justified because the number of observations in all experiments exceeds 30). We apply this statistical method to check the null-hypothesis that there is no statistically significant difference in the accuracies of models under consideration. Z is calculated according to the following formula (2):

$$Z = \frac{p_I - p_I I}{\sqrt{S_I^2 + S_I^2 I}} \qquad S = \sqrt{\frac{pq}{n}} \tag{2}$$

where p is the probability of correct classification of a review (i.e. accuracy); q – the probability of incorrect classification; n – number of observations; S – standard deviation; numbers I and II are the 2 models that are compared.

For a confidence level equal to 0.95 (α = 0.05) the null-hypothesis is confirmed if Z < 1.96.

Experiments

Testing one-parameter and two-parameter models

First, we aim to find out whether the Separate model outperforms the Joint model (1). In order to compare them we construct integral models over all the categories (Table 7) and the 'best' category models (Table 6), i.e. CARS (during the experiments it was noticed that the category CARS gives consistently better results than the rest of the categories).

Table 6: Accuracies of one-/ and two-parameter models built on CARS

	Model 1	Model 2	Model 3	Model 4	Model 5
1 variable	0.8	0.76	0.78	0.78	0.8
2 variables	0.8	0.68	0.72	0.8	0.78

Table 7: One-/ and two-parameter multicategorial models

	Model 1	Model 2	Model 3	Model 4	Model 5
1 var					
M:	4.71*score−0.69	4.18*score−0.73	3.08*score−0.72	3.55*score−0.73	1.88*score−0.68
A:	0,780	0,774	0,760	0,746	0,714
2 var					
M:	5.52*pos+3.73*neg −1.03	5.08*pos+3.08*neg −1.16	2.36*pos+3.9*neg−0.23	6.71*neg+1.15	3.76*pos−2.56
A:	0,783	0,774	0,746	0,703	0,706

Legend: M = Model; A = Accuracy

It is interesting to draw attention to the fact that when building regressions for two-parameter Model 5 the equation always transforms into a one-parameter one. This is due to the fact that for this model the overall sentiment contribution is equal to the number of sentiment-words. There is a functional dependency between the parameters pos and neg (neg = pos − 1), which makes regression impossible.

To compare the models we firstly apply the z-test (formula (2); n=400) to multicategorial models (Table 7), namely models 1, 2 and 3 as the other two have been transformed into single-parameter. The result is that for every pair Z<1.96; the null-hypothesis is confirmed, therefore, there is no statistically significant difference between the models. We then carry out the same test (formula (2); n=50) for the category CARS (Table 6) with the same result.

We conclude that in as far as there is no statistically significant difference, a one-parameter model is preferable.

Testing model granularity

Comparison of regression models for different granularity levels of sentiment words is accomplished using the one-parameter model as it proved to be preferable in the previous section. The constructed models and their accuracies for all individual categories are presented in Table 8 (A stands for model accuracy). We should note that regression could not be built on the category BOOKS due to a high level of inconsistency of sentiment-words contributions.

Table 8: One-parameter categorial models of different granularity scale

M:	books	cars	computers	cookware	hotels	movies	music	phones
1	–	8.37*X−1.07 A=0.8	5.77*X−0.85 A=0.8	4.4*X−0.71 A=0.74	5.54*X−1.1 A=0.8	4.48*X−0.49 A=0.76	4.54*X−0.58 A=0.72	3.79*X−0.61 A=0.7
2	–	6.55*X−1.02 A=0.76	5.06*X−0.87 A=0.82	4.18*X−0.78 A=0.74	5.08*X−1.16 A=0.82	4.03*X−0.54 A=0.74	3.9*X−0.6 A=0.72	3.31*X−0.62 A=0.7
3	–	4.61*X−0.99 A=0.78	3.85*X−0.88 A=0.76	3.44*X−0.85 A=0.76	3.95*X−1.18 A=0.78	2.98*X−0.53 A=0.76	2.7*X−0.56 A=0.7	2.25*X−0.6 A=0.7
4	–	5.58*X−1.1 A=0.78	4.13*X−0.89 A=0.72	3.64*X−0.79 A=0.7	4.46*X−1.1 A=0.78	3.49*X−0.55 A=0.74	3.27*X−0.58 A=0.7	2.69*X−0.59 A=0.68
5	–	2.91*X−1.04 A=0.8	2.29*X−0.86 A=0.7	2.27*X−0.83 A=0.74	2.48*X−1.08 A=0.74	1.97*X−0.53 A=0.72	1.51*X−0.47 A=0.62	1.19*X−0.49 A=0.6

Legend: M = Model

The models are compared using the z-test (formula (2)) where n = 50 for categorial models and n=400 for the integral model. The comparison is carried out in pairs: Model 1 is compared to all others. In Table 9 we present the results of the comparison of Model 1 to the one with the lowest accuracy within a category. Table 10 shows the comparison of multicategorial Model 1 to other multicategorial models. In as far as the maximum value of z-statistics for all categories (except for multicategorial Model 5) is less than Z-criterial, we infer that there is no significant difference between any compared models. Figure 3 presents a comparison of accuracy of models of different granularity within one-parameter regression for the best and the worst category.

Despite the fact that a change of the granularity scale does not give a statistically significant difference in the performance, it cannot escape our attention that the accuracy monotonically decreases with the roughening of the granularity scale (when an integral model is considered). Therefore, we do not exclude the possibility that a greater amount of data will reveal the higher impact of fine-grained sentiment scales.

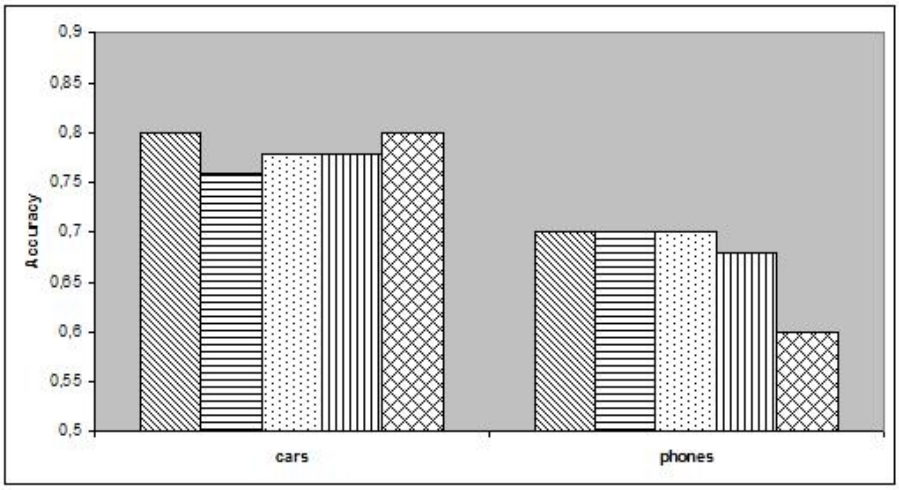
Table 9: Max z-statistics for each category

	cars	computers	cookware	hotels	movies	music	phones
Z-statistics	0.48	1.16	0.45	0.71	0.46	1.07	1.05
Z-criterial (5%)	1.96	1.96	1.96	1.96	1.96	1.96	1.96

Table 10: Z-statistics for multicategorial model

	Model 2	Model 3	Model 4	Model 5
Z-statistics	0.204	0.672	1.132	2.153
Z-criterial (5%)	1.960	1.960	1.960	1.960

Figure 3: Comparison in accuracy of five one-parameter models for CARS and PHONES



Testing the multicategorial model

For the purpose of studying the possibility of domain transfer and application of regression models to a corpus of unknown subject we apply one-parameter multicategorial models (Table 7) to each object category comparing the accuracy to that of corresponding categorial models. In Table 11 there are accuracies of local categorial models and of applied multicategorial models. Figures 4 and 5 show the comparison of accuracy for multicategorial and local models of type 1 and 5 for all categories.

To examine the proximity of the multicategorial model to categorial ones confidence intervals for accuracy in each category are calculated:

$$P = p \pm Z_\alpha \sigma \tag{3}$$

where P is a range of possible accuracy values; $Z_\alpha/2$ - quartile measure (α=5%, Z=1.96); σ – mean square value of accuracy. The results for Model 1 are presented in Table 12. The accuracy of every multicategorial model lies within the confidence interval of accuracy of categorial models; the multicategorial model is therefore representative for every object category.

Table 11: Accuracy of multicategorial models for each category

	Model 1	Model 2	Model 3	Model 4	Model 5
books_all	0.56	0.60	0.56	0.56	0.56
books_local	–	–	–	–	–
cars_all	0.86	0.80	0.80	0.82	0.80
cars_local	0.80	0.76	0.78	0.78	0.80
comp_all	0.84	0.82	0.80	0.74	0.70
comp_local	0.80	0.82	0.78	0.72	0.70
cook_all	0.78	0.76	0.78	0.78	0.76
cook_local	0.74	0.74	0.76	0.70	0.74
hotels_all	0.78	0.78	0.76	0.78	0.72
hotels_local	0.80	0.82	0.78	0.78	0.74
movies_all	0.78	0.80	0.82	0.76	0.76
movies_local	0.76	0.74	0.76	0.74	0.72
music_all	0.68	0.72	0.68	0.68	0.64
music_local	0.72	0.72	0.70	0.70	0.62
phones_all	0.74	0.72	0.70	0.68	0.68
phones_local	0.70	0.70	0.70	0.68	0.60

Table 12: Confidence intervals for model accuracy (Model 1)

	P	Multicat. Accuracy
cars	0.80±0.162	0.86
comp	0.80±0.167	0.84
cook	0.74±0.174	0.78
hotels	0.80±0.162	0.78
movies	0.76±0.172	0.78
music	0.72±0.184	0.68
phones	0.70±0.187	0.74

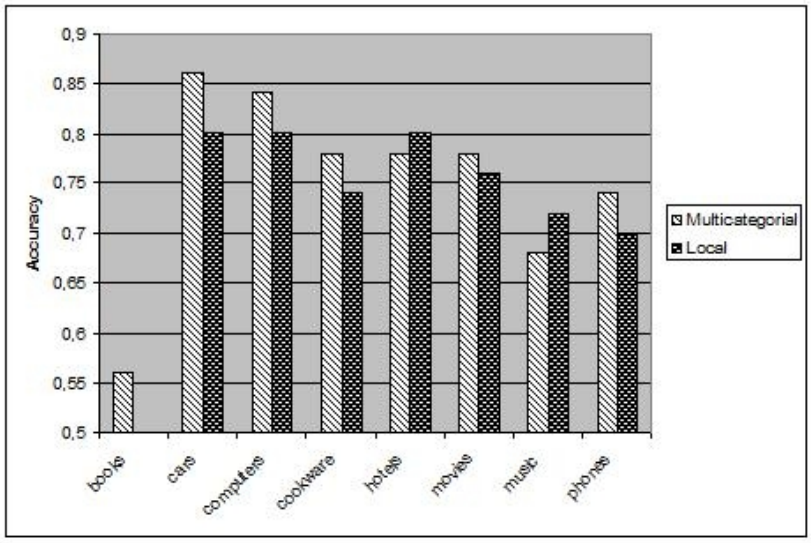
Conclusions

Discussion

In this study we have compared different regression models in the framework of binary sentiment classification (positive/negative). Applying the z-test to the results of leave-one-out cross-validation we firstly compared one-/ and two-parameter models, both local and multicategorial. The null-hypothesis was confirmed, which implies that there is no statistically significant difference between the given types of models. We therefore infer that in as far as there is no difference there is no reason to use a more complex model, namely the two-parameter one.

Comparison of five models of different granularity levels has shown the same result: there is no statistically significant difference between different types of categorial model (cf. Table 9). As for multicategorial models – the difference only appears at the level of model 5. The summary comparison of accuracy for the best and the worst category, Figure 3, showed that a model with a rougher granularity

Figure 4: Comparison of multicategorial and local models (for model 1)



scale can be used with no loss in performance.

The construction of the multicategorial model on the whole corpus produces a level of accuracy which correlates to the results obtained by Maite Taboada's research group (0,78 in our case vs. 0,83 – Brook, 2009). Note, that in contrast to Maite we did not take negation and intensification into account.

Applying a multicategorial model to local categorial ones (Table 14, Figures 4 and 5) showed that the multicategorial model is representative for every object category and can be successfully applied even to 'bad' categories.

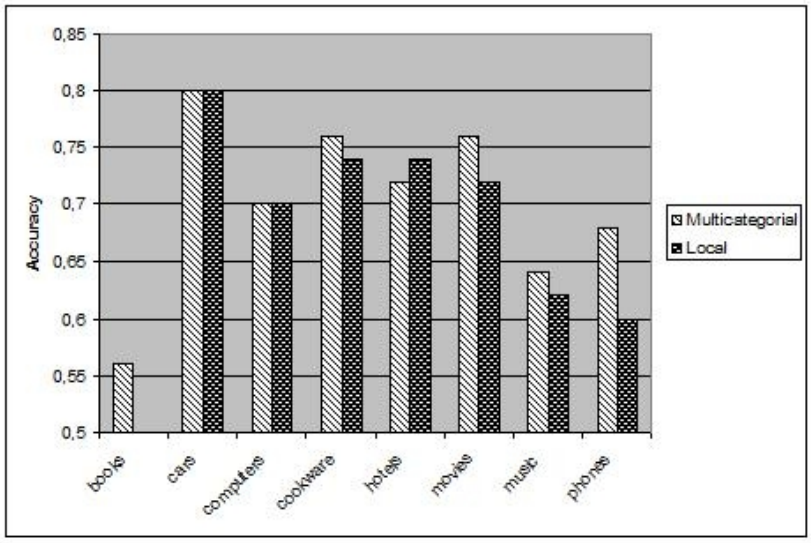
To sum up, we conclude that our results are not inferior to Maite Taboada's: simpler models can be used for the purpose of sentiment classification with no loss in performance as reducing the sentiment scale would in turn reduce the subjective influence of the raters (note that SO Dictionaries used in SO-CAL are manually rated sentiment-words and the finer the scale, the more difficult it is to assign correctly the value of the intensity of the emotion).

Future work

Possible developments of the present work might include construction of models for a more detailed scale of sentiment, that is, when reviews are classified not only within the binary polarity of positive vs. negative, but using a triple classification: positive, negative, neutral. The models for triple classification can be further analysed for the possibility of constructing models of 5 levels of sentiment classification: very positive, positive, neutral, negative, very negative.

The other option is testing the usage of Bayes classifiers for obtaining sentiment assessments and comparing the results with those when using regression

Figure 5: Comparison of multicategorial and local models (for model 5)



models.

In this study we tested our regression models on a corpus containing reviews obtained from the Internet, these reviews being written by ordinary users on product categories such as books, hotels, computers, etc. It would be useful to apply regression and Bayes models to specialised subject areas instead (economics, physics, etc), for example, with the aim of facilitating critical reviews of articles on a given area of knowledge.

References

Brooke, J., Tofiloski, M., Taboada, M. (2009): Cross-Linguistic Sentiment Analysis: From English to Spanish. In: *Proceedings of the 7th International Conference on Recent Advances in Natural Language Processing*. Borovets (Bulgaria), pp. 50–54.

Cramer, H. (1946): *Mathematical methods of statistics*. Cambridge, .

Hatzivassiloglou, V., McKeown, K. (1997): Predicting the semantic orientation of adjectives. In: *Proceedings of 35th Meeting of the Association for Computational Linguistics*. Madrid (Spain), pp. 174–181.

Hiroshi, K., Tetsuya, N., Hideo, W. (2004): Deeper sentiment analysis using machine translation technology. In: *Proceedings of the 20th international conference on computational linguistics (COLING)*. Geneva (Switzerland), pp. 494–500.

Kim, S.-M., Hovy, E. (2004): Determining the sentiment of opinions. In: *Proceedings of the 20th international conference on computational linguistics (COLING)*. Geneva (Switzerland), pp. 1367–1373.

Nasukawa, T., Yi, J. (2003): Sentiment analysis: capturing favorability using natural language processing. In: *Proceedings of the 2nd international conference on knowledge capture*. Florida (USA), pp. 70–77.

Pang, B., Lee, L. (2004): A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In: *Proceedings of the 42nd annual meeting of the Association for Computational Linguistics (ACL)*. Barcelona (Spain), pp. 271–278.

Pang, B., Lee, L. (2008): Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, Vol. 2, No. 1–2, pp. 1–135.

Pang, B., Lee, L., Vaithyanathan, S. (2002): Thumbs up? Sentiment classification using Machine

Learning techniques. In: *Proceedings of Conference on Empirical Methods in NLP* , pp. 79–86.
Taboada, M., Anthony, C., Voll K. (2006): Creating semantic orientation dictionaries. In: *Proceedings of 5th International Conference on Language Resources and Evaluation (LREC)*. Genoa (Italy), pp. 427–432.
Taboada, M., Grieve, J. (2004): Analyzing appraisal automatically. In: Qu, Y., Shanahan, J. G., Wiebe, J. (eds.) *Proceedings of AAAI Spring Symposium on Exploring Attitude and Affect in Text (AAAI Technical Report SS-04-07)*. Stanford University, CA : AAAI Press, pp. 158–161.
Turney, P. (2002): Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In: *Proceedings of 40th Meeting of the Association for Computational Linguistics* , pp. 417–424.

Yi, J., Nasukawa, T., Niblack,W., Bunescu, R. (2003): Sentiment analyzer: extracting sentiments about a given topic using natural language processing techniques. In: *Proceedings of the 3rd IEEE international conference on data mining (ICDM)*. Florida (USA), pp. 427–434.
Wilson, T., Wiebe, J., Hoffmann, P. (2005): Recognizing contextual polarity in phrase-level sentiment analysis. In: *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*. Vancouver (B.C., Canada), pp. 347–354.

We would like to thank Maite Taboada for kindly providing SO-Dictionaries and a review corpus for the present study.

A Look at Wikipedia Readability: Language, Domain and Style

Olga Ogurtsova
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: assailable@yandex.ru

Mikhail Alexandrov
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: malexandrov@mail.ru

Xavier Blanco
Department of French and Romance Philology, Faculty of Philosophy and Arts, Autonomous University of Barcelona, 08193 Bellaterra, Spain
e-mail: Xavier.Blanco@uab.cat

Biographical note

Olga Ogurtsova completed a secondary school with a linguistic bias and she graduated from the Russian State Pedagogical University of Herzen in 2008. Her speciality is teaching methods of the Spanish and English languages. At the moment she is a student on a Master's program of the Autonomous University of Barcelona in Spain. The area of her scientific interests is linguistic aspects of internet technologies.

Mikhail Alexandrov is a member of the LexSem Research group of the Autonomous University of Barcelona in Spain. He is a professor of the Academy of National Economy under Russian Government. He is an applied mathematician and author of numerous publications related to mathematical modelling and natural language processing. His current topics of research are machine learning (inductive modelling, clustering) and internet technologies (social networking).

Xavier Blanco is full professor at the Autonomous University of Barcelona (UAB) in Spain. He is author of several large-coverage electronic dictionaries of Spanish for machine translation software and other NLP applications. He is also author of numerous scientific and technical papers concerning translation studies, lexicology, phraseology and lexicography. He coordinates the International Master's program in Natural Language Processing and Human Language Technology in the UAB.

Introduction

State-of-the-art

Wikipedia 'is a multilingual, web-based, free-content encyclopaedia project based on an openly-editable model' (http://en.wikipedia.org/wiki/Wikipedia:About). Wikipedia is written collaboratively by largely anonymous Internet volunteers who write without being paid. Wikipedia content is intended to be factual, notable, verifiable with cited external sources, and neutrally presented. The basic principles of Wikipedia's Arbitration system and committee (known colloquially as 'Arbcom') were developed mostly by Florence Devouard, Fred Bauder and other key early Wikipedians in 2003. The principles can be found in Wikipedia itself (http://en.wikipedia.org/wiki/Wikipedia:About).

Over the years, interest in the Wikipedia phenomenon has been growing. Probably it reached its peak in 2008. At present there is a huge number of works that deal with Wikipedia, and conferences devoted to the Wiki-like resources are held in different cities. There are investigations into the accuracy of Wikipedia, collaborative work of the authors of Wikipedia articles, vandalism in Wikipedia, content of the different language sections of Wikipedia (Giles, 2005; Shapovalov and Malutina, 2009; Belani, 2009; Biuk-Aghai and Lei, 2010), etc.

The homogeneity of Wikipedia style is one the principal requirements for Wikipedia texts. This concerns the homogeneity both of texts that belong to different domains and different languages. Readability is the easiest stylistic characteristic of text to compute, which is why it became the subject of our research.

Readability is defined as the level of complexity of text comprehension, which is determined by certain computable linguistic and stylistic features, such as average lengths of sentences and words, average frequency of rare words and average number of prepositions in phrases, etc. Various indexes are used to evaluate the readability/complexity of a text: the Dale-Chall Readability Formula, Flesch readability index, Farr-Jenkins-Patterson Formula, Fry Readability Formula, Fog Index, Lorge formula, and SMOG Grading (Dale and Chall, 1948; Flesch, 1948; Farr, Jenkins and Patterson, 1951; Fry, 1977; Gunning, 1952; Lorge, 1939; McLaughlin, 1969).

Investigations showed that readability depends on genre (novels, newspapers, scientific papers, etc) and this problem has been considered in publications related to the mentioned indexes. The readability of textbooks for schoolchildren and students are the most referenced topic in index descriptions and their applications (`http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test`).

The most popular readability index among western researchers is the Flesch index. It uses average sentence length measured in words and average word length measured in syllables: the bigger this index, the higher the ease of reading. There are programs designed for calculation of this index, for example Word Counter for Macintosh OS X or INFLESZ for Windows 9x. y NT/XP (`http://www.legibilidad.com/home/acercade.html`). Some Internet applications contain functions that permit calculation the Flesch index. For example, the Flesch index for different languages can be calculated on the web-page (`http://www.standards-schmandards.com/exhibits/rix/`). There is a program for calculation of the Flesch index for Russian and English text (Oborneva, 2005) but this program is not a free-share one.

The Mikk index is known among Russian researchers (Tuldava, 1975; Makagonov 1998). Mikk's formula includes the same variables as Flesch's formula, namely the average sentence length and the average word length in a text. However, this index reflects the complexity and not the readability of a text. In other words the higher the index is the more difficult a given text is to read. The bibliography lacks references to software that counts the Mikk index. We used the program TextComplexity developed in the department of French philology of the AUB. This program was used also by one of the authors when she was working on her Master's thesis (Ogurtsova, 2010).

It is worth mentioning that we did not find publications where the problem of stylistic homogeneity of Wikipedia texts had been considered. The proximity

of the Wikipedia style to scientific style has not been evaluated as well.

Problem settings

There are two problems to be considered in this article.

1. We want to check whether the administrators of Wikipedia adhere to equal readability requirements for texts in different languages and domains. We have chosen two domains – physics and linguistics, which contrast as natural and humanitarian disciplines. That is why a comparison of texts from these domains would reflect the situation in other domains less contrasted in their content. As concerns the languages, we have chosen English and Spanish, which do not belong to the same language group (as Romance, Slavic or Finno-Ugric languages). The Flesch and Mikk indexes are used to compare texts. The problem consists in the calculation of the mean value of the indexes for groups of texts and testing a hypothesis about the statistical significance or non-significance of the differences.
2. We want to know to which sub-style of the scientific style (properly scientific, popular or didactic) texts from Wikipedia belong. English texts on linguistics representing the three mentioned sub-styles were selected in order to be compared with texts from Wikipedia. In this case we use the Mikk index only for the comparison. As well as in the previous case we need to find the mean of the index for each group of texts and then to test the statistical significance or non-significance of the differences.

The paper contains 4 sections. The next section describes the method of investigation. This method consists of testing the hypothesis about non-significant differences in the readability indexes mentioned above. Section 3 presents the results of experiments. The discussion is included in section 4.

Decision making

Modified Flesch and Mikk indexes

Main formulae

The following Flesch formula is accepted for the English language (Flesch, 1948):

$$\mathrm{IF} = 206.84\text{-}84.6S - -1,015M$$

where S is the average length of a word in syllables and M is the average length of a sentence in words.

The following Flesch formula is accepted for the Spanish language (`http://www.legibilidad.com/home/acercade.html`):

$$\mathrm{IF} = 206.84\text{-}62.3S\text{-}M.$$

Table 1 demonstrates the correspondence between the Flesch index and level of readability

This is the Mikk formula:

$$IM = SLn(M)$$

where S is the average length of a word in syllables and M is the average length of a sentence in words.

The equation of regression for word length in characters and syllables

Revealing syllables in words is a procedure that needs to take into account the linguistic properties of a given language. Unfortunately we did not find in the literature any universal free-share software that could do it. In this situation it is reasonable to consider the possibility to substitute syllables with characters. But such a substitution needs justification.

For this we selected pieces of text from several English documents (newspaper) and took the 100 most frequent words. Then we calculated the number of characters and syllables in each word and constructed a regression

$$y = 2.86x \tag{1}$$

where y is the quantity of characters and x is the quantity of syllables in a word. The coefficient of correlation between the mentioned values is equal to 97,4%.

The same experiment was done with words from the Spanish language. The following dependence was revealed:

$$y = 2.35x. \tag{2}$$

Here the coefficient of correlation is equal to 98,8%.

All the computations were completed with the MegaStat package.

Modification of the Flesch and Mikk formulae

The dependences (1) and (2) obtained were utilised to substitute the number of syllables with the number of characters in the Flesch formulae. Thereby, after substitution the following Flesch formulae were obtained.

For the English language:

$$[\mathrm{IF} = 206.84 - -84.6/2.86N - -1.015M.$$

Here, N is the average length of a word in characters and M is the average length of a sentence in words.

For the Spanish language:

Table 1: Flesch values and level of readability

Flesch index	Readability
70-80	very easy (novels)
60-65	normative (newspapers)
50-55	intellectual level (business editions, literary magazines)
30 and lower	scientific level (professional and scientific literature).

$$\mathrm{IF} = 206.84\text{-}62.3/2.35N\text{-}M.$$

The Mikk formula was modified formally; we only substituted the number of syllables with the number of characters without any transformation. This substitution is reasonable because it changes all the values of tests in the same proportion. So this formal modification has no an impact on the testing hypothesis. Therefore we have the following Mikk formula:

$$IM = SLn(M)$$

where N is the average length of a word in characters and M is the average length of a sentence in words.

One should note that such a substitution of syllables with characters is justified when we measure the average length of words in a whole text. It does not fit for cases when we have to analyse concrete words.

Comparison of indexes

In all our experiments we compare readability of two document sets using Flesch and/or Mikk indexes. If a given index (Flesch or Mikk) has close values for each set then one can say that these sets have close styles from the point of view of readability. The comparison of indexes is performed statistically in the framework of testing the hypothesis about non-significance in differences of index means.

For testing the hypothesis we use the standard technique of p-value (Cramer, 1999). It consists of two steps:

1. One calculates the means (m_1,m_2) and deviations of the means (s_1,s_2) of a given index for two data sets and then forms the so-called t-statistics

$$ts = |m2 - m1|/\sqrt{({s_1}^2 + {s_2}^2)}$$

2. One calculates the probability of the extreme case that random t-statistics reaches this value. $p = P(t > t_s)$

   The lower the p-value, the less likely the result is if the hypothesis is true. Let we fix a level of significance α (10%, 5%, 1%). This level defines the probability of error when we reject the true hypothesis. The technique of hypothesis testing consists in the following rule:

   Hypothesis is accepted if $p > \alpha$

   Hypothesis is rejected if $p \leq \alpha$

   When we reject the hypothesis we can make an error with the probability α (type 1 error)

There are standard functions in all popular packages related to experimental data processing, which calculate p-value for given t-statistics or for two given data sets. For example such functions are included in the list of standard Excel-functions.

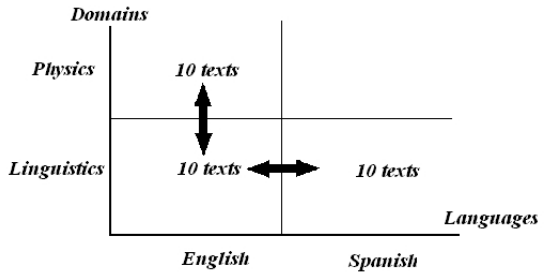


Figure 1:
Languages and domains used in the experiments

Note: If the number of data in each document set is more than 30 then one should use functions for working with the normal distribution. If the number of data is equal to or less than 30 then one should use functions for work with the Student distribution. In the latter case it is necessary to take into account the so-called degree of freedom for the Student distribution. In our case this value is equal to $k = 2n - 2$, where n is the number of documents in each document set.

Experiments

Analysis of the homogeneity of Wikipedia by language and domain

Plan of the experiments

In the first series of experiments we compared different texts from Wikipedia. In order to check the homogeneity of Wikipedia by language we need to take texts from the same domain written in two different languages. In order to check the homogeneity of Wikipedia by domain we need to take texts from two different domains in the same language. In such a way we can essentially reduce the number of experiments and simplify the interpretation of results.

We consider:

- contrasting domains, i.e. linguistics and physics;
- languages from different groups, i.e. English and Spanish.

The plan of the experiments in the co-ordinates language-domain is represented in Figure 1.

The Flesch and Mikk indexes for English and Spanish texts

According to the plan presented in Figure 1 we examined 20 texts on linguistics from the English and Spanish versions of Wikipedia. The average text length was approximately 1,300 words both for English and Spanish documents. The means and variations of the Flesch and Mikk indexes were calculated for each set. Table 2 contains the results of the calculations.

Our goal is to test the null-hypothesis about non-significance of differences in the means for each index. We use here the standard procedure described in section 2.2. The source data is given in Table 2. The degree of freedom is equal to

Table 2: Means and deviations for Flesch and Mikk indexes

Value	Flesch English	Flesch Spanish	Mikk English	Mikk Spanish
Mean	23.62	39.24	16.72	17.34
Deviation	7.59	7.71	1.20	1.29

Table 3: Verification of the null-hypothesis about the non-significance of differences

p-value	p-critical	Result	Flesch index
0.0002	1%	The difference is significant	Mikk index
0.28	1%	The difference is non-significant	

$k = 2n - 2 = 18$, importance level αy = 0.01. Table 3 contains the p-value for each test and the result.

Therefore the Flesch index shows the significant difference for texts in English and Spanish, while the Mikk index does not detect this difference. This fact can be explained by the influence of natural differences in the two languages (English words are shorter and English sentences are longer then Spanish ones) which is not considered by the Mikk index.

The Flesch and Mikk indexes for texts on linguistics and physics

According to the plan presented in Figure 1 we examined 20 English texts on linguistics and physics. The average text length was equal to approximately 1,300 words for documents on linguistics and 1,200 words for documents on physics. The means and variations of the Flesch and Mikk indexes were calculated for each set. Table 4 contains the results of the calculations.

Table 4: Means and deviations for Flesch and Mikk indexes

Value	Flesch Linguistics	Flesch Physics	Mikk Linguistics	Mikk Physics
Mean	23.62	33.41	16.72	15.68
Deviation	7.59	6.36	1.20	0.75

To test the null-hypothesis we complete the same procedure described in the previous item. Table 5 contains the p-value for each test and the result.

Both Flesch and Mikk indexes revealed an absence of significant differences between the texts on physics and linguistics.

The graphical illustration of the results obtained relative to the average values is presented in figures 2 and 3.

Scientific style of Wikipedia

Functional styles

Functional style is a variety of the literary language performing a specified function in communication (Solganik, 1997). In this paper we examine the scientific style and its sub-styles: the properly scientific, popular and didactic.

Table 5: Verification of the null-hypothesis about the non-significance of differences

p-value	p-critical	Result	Flesch index
0.01	1%	The difference is non-significant	Mikk index
0.03	1%	The difference is non-significant	

Figure 2: Flesch indexes for the texts

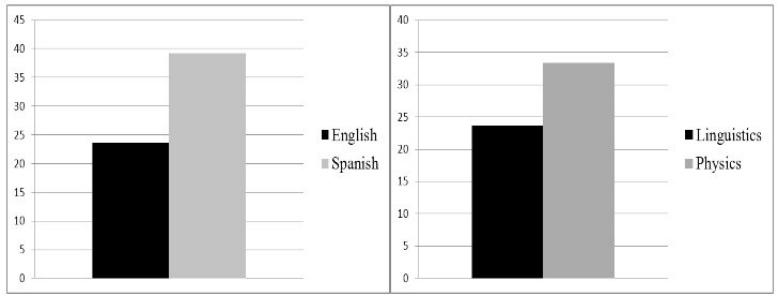
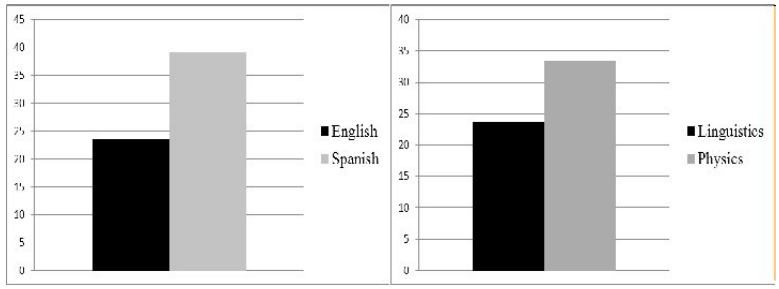


Figure 3: Mikk indexes for the texts



Scientific style is characterised by the logical sequence of phrases and it is characterised by accuracy, brevity and absence of ambiguity. The research paper is forwarded to a reader specialising in a concrete scientific branch and possessing knowledge at about the same level as the author of article, the sender.

The specificity is inherent also in the popular sub-style. The receiver of such articles is a person interested in this or that science. Within the limits of this sub-style some deviations from norms are supposed – the use of words in their figurative sense is possible.

Documents with a didactic sub-style are addressed to future specialists, students and schoolchildren. Its purpose is to teach and to describe the facts that are necessary for the acquisition of material.

We consider the degree of correspondence of Wikipedia documents to the above mentioned sub-styles of the scientific style from the point of view of text complexity. Text complexity is evaluated here with the Mikk index.

Certainly, scientific style and its sub-styles are not restricted only by text complexity. The full characteristic of the style requires testing other formally-statistical and informally-linguistic indicators. For example, harmony of texts and richness of text vocabulary refer to the first one, while specific idioms refer to the second. However, in the present work we limit our consideration to text complexity only.

In our experiments we used documents on linguistics in English. We selected 10 texts from Wikipedia, 10 scientific articles, 10 popular articles and 10 manuals. Therefore we had 40 texts in total. The sources of the scientific articles are various theses on linguistics and research described in articles for scientific journals. The popular texts were taken from electronic newspapers or journals for a wide range of readers and from encyclopaedia articles. The didactic materials are more varied. We were working with fragments from textbooks on linguistics for high school students and for students of the first years of university (both from faculties of languages and from technical faculties).

The means and deviations of the Mikk index were calculated for each group of texts. The results are presented in Table 6.

Table 6: Values of the Mikk index for texts on linguistics

Category	Mean	Variation
Wikipedia	16.72	1.2
Scientific articles	17.04	1.1
Popular articles	15.88	1.35
Didactic materials	12.67	2.44

The means of the Mikk index are presented on the Figure 4.

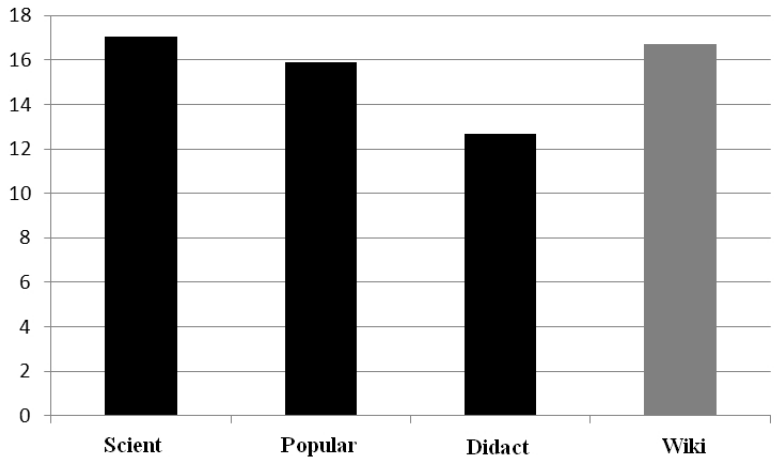
We tested the significance of differences in Mikk index between the Wikipedia articles and the other three groups of texts. We verified the null-hypothesis about the non-significance of differences of the means. The verification of the hypothesis was done as described in part 'Comparison of indexes'. The results are presented in Table 7.

Table 7: Verification of the zero-hypothesis about the non-significance of differences

	p-value	p-critical	Result
Scientific articles	0.54	1%	The difference is non-significant
Popular articles	0.16	1%	The difference is non-significant
Didactic materials	0.0004	1%	The difference is significant

Therefore the level of complexity of Wikipedia texts differs significantly from the level of complexity of didactic texts. It is close to the level of complexity of scientific and popular texts.

Figure 4: Values of the Mikk index for different text categories



Conclusion

Discussion

- We revealed the dependences between average number of characters and syllables in English and Spanish words. This relation has a high correlation (97%-99%). Based on this relation we could modify the Flesch formula and justify the possibility of using the Mikk formula with a formal substitution.
- A significant difference in the Flesh index for English and Spanish texts was shown, although the Mikk index did not detect such a difference. In the experiments, documents on linguistics were used.
- It was shown that there were no significant differences in Flesh and Mikk indexes for linguistics and physics. In the experiments, English documents were used.
- The level of text complexity for Wikipedia articles is close to the text complexity of scientific and popular documents and differs only from manuals. In the experiments, documents on linguistics were used.

Our conclusions were based on very limited document sets. So, the obtained results can be considered only as preliminary ones, which should be tested once more on a larger corpus of documents.

Future work

In the future, we consider:

- repeating the completed experiments on large data sets containing dozens and even hundreds of documents;

- enlarging the number of languages, in particular, to consider texts in the Russian and German languages;
- enlarging the number of domains, in particular, to consider economics and history;
- considering other formally-statistical indicators of style, such as the degree of lexical richness of text.

References

Wikipedia:About [on-line]. (c2010): [cit. 2010]. Available at: `http://en.wikipedia.org/wiki/Wikipedia:About`.

Belani, A. (2009): *Vandalism Detection in Wikipedia: a Bag-of-Words Classifier Approach Master's*. Cornell University„ .

Biuk-Aghai, R. P.; Lei, Keng Hong (2010): Chatting in the Wiki: Synchronous-Asynchronous Integration.. In: *Proceedings of the 6th International Symposium on Wikis and Open Collaboration*.. Gdańsk (Poland) .

Cramer, H. (1999): *Mathematical methods of statistics*. Princeton University Press„ .

Dale, E., Chall, J. S. (1948): A formula for predicting readability.. *Educational research bulletin*, Jan. 21 and Feb. 17, 27, pp. 1–20, pp. 37–54.

Flesch, R. (1948): A new readability yardstick.. *Journal of Applied Psychology*, Vol. 32, pp. 221–233.

Farr, J. N., Jenkins J. J., Paterson D. G. (1951): Simplification of the Flesch Reading Ease Formula.. *Journal of Applied Psychology*, Vol. 35, No. 5, pp. 333–357.

Flesch-Kincaid readability test [on-line]. (c2010): [cit. 2010]. Available at: `http://en.wikipedia.org/wiki/Flesch-Kincaid_readability_test`.

Fry, E. (1977): *Elementary Reading Instruction*. New York, .

Giles, J. (2005): Internet encyclopedias go head to head.. *Nature*, December 15, pp. 900–901.

Gunning, R. (1952): *The technique of clear writing*. New York : McGraw-Hill, .

...la legibilidad? (in Spanish) [on-line]. (c2010): [cit. 2010]. Available at: `http://www.legibilidad.com/home/acercade.html`.

Lorge, I. (1939): Predicting reading difficulty of selections for children.. *Elementary English Review*, 16, pp. 229–233.

Machine Learning [on-line]. (c2010): [cit. 2010]. Available at: `www.machinelearning.ru`.

Makagonov, P., Alexandrov, M. (1998): Analysis of contents of educational courses with special statistical and graphical methods.. In: *Proc. XV World Computer Congress, Conf. "Téleteaching '98"*.. Vienna : Austrian Comp. Soc., pp. 679–683.

McLaughlin, G. Harry (1969): SMOG Grading – a New Readability Formula.. *Journal of Reading*, Vol. 12, No. 8, pp. 639–646.

Oborneva, I. V. (2005): Mathematical model for evaluation of didactic texts.. *Proc. of Moscow State Pedag. Univ., series 'Informatics'*, Vol. 4, No. 1, pp. 141–147.

Ogurtsova, O. (2010): The statistical index of readability as a formal indicator of a style of scientific/popular/didactic texts. .

Readability index calculator [on-line]. (c2010): [cit. 2010]. Available at: `http://www.standards-schmandards.com/exhibits/rix/`.

Shapovalov, R., Malutina, A. (2009): *About the mission of language sections of Wikipedia different from the English Wiki-conference 2009*. Saint-Petersburg, .

Solganik, G. Y. (1997): *Stylistic of text: Manual*. Moscow, .

Tuldava, Yu (1975): About Measurement of Text Difficulties.. In: *Proc. Of Tartu State University*.. , pp. 102–120.

The authors thank Prof. Julio Murillo from the Department of French and Romance Philology of the Autonomous University of Barcelona for his valuable advice and consultations.

Data Mining of Online Judicial Records of the Networked US Federal Courts

Joseph Zernik
Human Rights Alert (NGO), PO Box 526, 91750 La Verne, California, U.S.A.
e-mail: jz12345@earthlink.net

Abstract

The US federal courts have completed a decade-long project of networking coast to coast. Data mining was conducted through the online public access system to examine the validity and integrity of records and of the system as a whole. Many records were not verified at all. Moreover, records were universally missing their authentication counterparts, required by law to render them valid and effectual. The authentication counterparts—previously public records—were now excluded from public access. Records, which are today posted online in the public access system, included both valid and invalid, void records. However, the public was unable to discern the difference. The system as a whole was deemed invalid. Case management systems of the courts must be subjected to certified, functional logic verification. Mandated system transparency should permit ongoing data mining, and the computing/informatics community should lead the way in monitoring the integrity of the courts in the digital era.

Key words

relational databases; functional logic verification; Case Management Systems; United States Courts; human rights; Los Angeles; California; United States; justice system; law; fraud; corruption; judges

Biographical note

Professor Zernik served on the faculty of the University of Connecticut, University of Southern California, and University of California, Los Angeles.

In 2010 he founded Human Rights Alert (NGO), dedicated to discovering, archiving, and disseminating evidence of human rights violations by the justice systems of the State of California and the United States in Los Angeles, California, and beyond. Special emphasis is given to the unique role of computerized case management systems in the precipitous deterioration of the integrity of the justice system.

Paper v Digital Administration of the Courts[i]

The transition of the US courts to digital administration was executed over a comparatively short time through a large-scale project managed by the Administrative Office of the US Courts, an arm of the US judicial branch. Dual systems were established: PACER—for public access to court records, and CM/ECF—for case management/electronic court filing. The systems are effectively a series of relational databases.[ii] With it—a sea change was affected in court procedures. In contrast,

procedures for the paper administration of the English-speaking courts evolved over centuries, and formed the foundation of due process and fair hearings rights. Disambiguation of court procedures and court records was the cornerstone of such rights. Therefore the current report employed data mining to examine the newly established digital administration of the US courts for the following fundamentals of due process: a) valid, published rules of court, b) public access to judicial records—to inspect and to copy, and c) verification 1 and authentication of judicial records, including notice and service of judicial papers. Finally—validity of the systems as a whole as assessed.

Valid Published Rules of Court

Procedures of the US courts under paper administration evolved from the English legal system, and were established in the US Code of Civil/ Criminal Procedures[iii, iv] and published Local Rules of Court, which the courts were authorized to adopt, [v] subject to prior publication of such rules for public comment and challenge. The transition to digital administration of the US courts inevitably entailed a sea change in court procedures, which had to be established by law or by the publication of new local rules of courts. The current report documents the failure of the US courts to publish their new rules pertaining to the new digital procedures.

Public Access to Judicial Records—to Inspect and to Copy

The right to public access to judicial records—to inspect and to copy - was well established in both US and common law.[vi] Judicial paper records, which were maintained by the Clerk of the Court, included individual Court Files together with Books of Courts, including, but not limited to the Court Dockets—logs of all valid proceedings and all valid records pertaining to the respective court files by the clerk. The transition to digital administration of the courts entailed substantial changes, not by necessity, of the well-established set of judicial records. Moreover, through differential individual authorities, it became much easier to conceal digital judicial records. The current report documents the universal denial of public access to critical judicial records, that is the authentication counterparts of individual court records in all courts that were examined.

Verification and Authentication of Judicial Records

Verification by a judge, and authentication by a clerk, of court orders and judgments were a prerequisite for entry of court records as honest, valid, and effectual court papers in court files and dockets. These requirements were founded in the US Constitution, in early Acts of US Congress, and in the US Code of Civil/ Criminal Procedures.[vii, viii]. Notice and service of judicial records—court minutes, orders, and judgments were an integral part of the authentication procedures mandated on the clerks for all court minutes, orders, and judgments. The clerks would mail out to all parties in a case certified, authenticated copies of all judicial records, jointly satisfying both the authentication per se and the notice and service requirements. In the transition to digital administration of the courts, new procedures were devised for the digital verification by judges, for the purported certification of entry

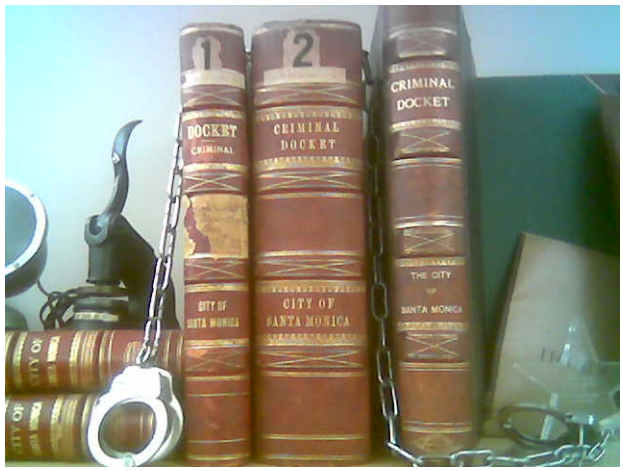


Figure 1:
Historic, paper-based Books of Court–Criminal Dockets; City of Santa Monica, California

of the records, and for the construction of dockets by the clerks. Additionally, service and notice were made possible through electronic mail. The current report documents deficient or invalid authentication of court records, effectively eliminating any valid certification by the clerk, and universal denial of public access to the authentication records.

Proposed Solutions

The current report documents the compromised integrity of the US courts, undermined through the installation of invalid, unverified digital administration systems. Solutions are readily available tha which could make digital administration of the courts honest and secure and far superior to paper administration of the courts. The relational databases,[ix] which form the foundation of digital administration of the US courts, must be subjected to certified, functional logic verification. Transparency should be required, to allow ongoing data mining by the public at large–for monitoring integrity of the courts. The computing/informatics community should assume a leading role in the safeguarding of human rights and the democratic nature of society in the digital era.

Objective

The current report investigated through data mining the online public records of the US courts and evaluated the safeguards for the fundamentals of due process and fair hearings in the transition from paper to digital administration of the courts. Moreover, the report examined the potential role of data mining of online public records of the courts as an essential civic duty–public monitoring of the integrity of the courts.

Figure 2:
CM/ECF—the Case Management/Electronic Case Filing system of the United States courts

The Systems

PACER,[x] the online system for Public Access to Court Electronic Records of the US Courts, permits public access on payment to records of the US district courts and some US courts of appeals, including the indices, dockets, calendars, and records of the various cases. CM/ECF[xi], the online system for Case Management/Electronic Court Filing of the US Courts, permits the courts themselves and attorneys, who are authorized by a given US court in a given case, to file records. Pro se litigants—those representing themselves in court, including most prisoners[xii] and others, who petition the US courts for protection of Human, Constitutional, and Civil Rights, are routinely denied access to CM/ECF, and are required to file records on paper. They are likewise permitted access to court records through PACER alone. CM/ECF also provides notice and service of records, only to those authorized, via email.

NEFs (Notices of Electronic Filing)[xiii]—at the US district courts, and NDAs (Notices of Docket Activity)—at the US courts of appeals, were established by the US courts as the counterparts for the purported digital authentication of court records.[xiv] Parties, who are unauthorized in CM/ECF are routinely excluded from notice and service of the NEFs and NDAs.

Methods—Data Mining and Record Examination

Data mining, which formed the basis for the current report, was conducted through public access to PACER, as permitted by law, in over 20 US district courts, US bankruptcy courts, and US courts of appeals.[xv] Some of the individual cases that were examined as part of the study were identified through methods which were developed for rapid screening of cases in the various courts to identify cases that were deemed at high risk of perversion and were therefore selected for further examination. Other cases were identified through direct alerts by individuals who were parties to the cases. Most of the cases involved litigations where individuals who were plaintiffs filed complaints for protection of civil rights, or complaints alleging wrongdoing by large financial institutions. Court records were examined in the individual cases to determine whether court minutes, orders, and judgments were verified and authenticated in a valid manner.[xvi] Since public access was found to be universally denied to the authentication counterparts, in some of the cases repeat attempts were made over months, as described below, to gain access to the authentication records, through written requests as well as repeat personal appearances at the offices of the clerks of some of the courts. In particular instances where

credible evidence was found that records provided through PACER were false and misleading, lacking verification and/or authentication, but displayed as "entered", written requests were filed with the clerks and presiding judges of the US district courts to investigate such records and initiate corrective actions.

Additionally, local rules of courts, general orders, and CM/ECF user's manuals were downloaded through routine web browsing of the web pages of the various courts. Court rules, general orders, and user's manuals were examined to determine whether the courts published clear and unambiguous rules establishing the courts' new digital procedures. When no clear and unambiguous published rules were found, requests were forwarded in some cases to the clerks and the chief judges to disclose the new rules of the courts.

Results

Data Mining

All data mining in the current report was manually performed. However, although not detailed in this report, methods were devised to scan large numbers of cases in any given US court, which could be automated. Although significant differences were found in implementation of PACER and CM/ECF in the various courts, the basic platforms were identical in all courts that were examined. The basic findings were likewise similar in all courts. Specific examples are provided below.

Published rules of court

Review of the local rules of court of the various US courts and US courts of appeals,[xvii, xviii] universally revealed no direct reference to PACER, CM/ECF, or new digital procedures pertaining to verification and authentication of, notice and service of, and public access to, court records. In some cases reference, albeit vague and ambiguous, was found in general orders of the courts. However, in the vast majority of the courts, which were examined, no reference was made anywhere in local rules of court, in general orders, or even in user's manuals, to the critical new procedures now established by the courts for purported authentication of records through a "Document Stamp" (an encrypted checksum string) as the equivalent of the former signature by the clerk of the court.

For example, upon inquiry, the office of the clerk, US District Court, Central District of California, claimed that the new rules were established in General Order 08-02.[xix, xx] In pertinent parts,[xxi] General Order 08-02 prohibited external hyperlinks in any court records, declared that acceptance of electronic filing constituted entry of a record in the case docket, established the NEF as the certification by clerk of court records, and the electronic "document stamp" as validating the authenticity and notice of the record. The order established that pro se litigants would be required to continue filing records on paper, and electronic filing of records, which were filed by pro se litigants on paper, would be executed by the clerk. However, the NEF itself was never defined in the General Order 08-02. Moreover, it should also be noted that General Order 08-02, in contrast to other general orders of the

US District Court, Central District of California, was published with no verification by a judge at all, and with no name of its author either. No significant information pertaining to electronic filing practices were found in the Local Rules of Court or General Orders of the US Court of Appeals, 9th Circuit.[xxii]

Additional critical information regarding the NEF and procedures of the US District Court, Central District of California was detailed in the Unofficial Anderson Manual.[xxiii] The CM/ECF User's Manual of the US Court of Appeals, 9th Circuit,[xxiv] likewise included detailed description of the NDAs—Notices of Docket Activity—as the purported authentication records of the Court.

Public Access to Judicial Records—to Inspect and to Copy

Public access to the purported authentication records—NEFs and NADs—was universally denied in PACER. Additional efforts were made to access the NEFs and NDAs in some of the courts, through repeated direct requests to the clerks of the courts. The courts routinely denied such access, with no reason at all. Limited access was eventually gained in two of the courts – the US District Court for the Central District of California, and the US Court of Appeals, 9th Circuit.[xxv]

It should further be noted that key records were removed from public access in the dockets of the various US district courts and US courts of appeals, with no reason at all. For example, in the Fine v Sheriff Appeal (09-56073) at the US Court of Appeals, 9th Circuit, records Dkt #28, 32, 36, 56, were removed from public access. These records were mostly related to response to the appeal by the Appellees. Such selective denial of public access to court records was of particular concern, since the Appellant in the case objected to some of the same response papers, claiming that they included as evidence records that were not admissible. PACER, as a public access system, if validated, should never permit such selective removal of records from the dockets, unless by authorized personnel, pursuant to sealing orders, docketed in the respective cases.

It should also be noted that the dockets of the US Court of Appeals, 9th Circuit, as a rule failed to explicitly state within the public dockets the date of entry of judgments of the US district courts from which the appeals were purported to originate. The dates of entry of judgments are critical data regarding the validity of the appeals as a whole, since the jurisdiction in appeals is limited by law to entered judgments, if notice of appeal is filed within the time frames permitted by law.

Table 1: Summary of information pertaining to electronic filing for the US District Court, Central District of California and the US Court of Appeals, 9th Circuit

Court	Rules of Court	General Orders	User's Manuals
Central District of California	None	General Order 08-02	Detailed description of electronic filing practices in an "unofficial" manual
Court of Appeals, 9th Circuit	None	None	Detailed description of electronic filing practices.

Verification, Authentication, and Notice of Judicial Records

The dockets of the US courts and courts of appeals were found to include records that were either not verified by a judge, or not adequately authenticated by a clerk. For example, two critical records—the Judgment and the Mandate, in the docket of Richard Fine v Sheriff of Los Angeles County (2:09-cv-01914) at the US District Court, Central District of California, were deficient. One was an unverified record, failing to include any signature by a judge, and both were unauthenticated records—missing the electronic "Document Stamp" in the respective NEFs:[xxvi] Similarly - orders of the US District Court, Vermont, in Huminski v Rutland Police Department (1:99-cv-160) were served with NEFs bearing no "Document Stamp" at all.[xxvii] Therefore such records could not possibly be deemed by the courts themselves as honest, valid, and effectual judicial records.[xxviii] Regardless, these records were listed in the dockets of the respective US district courts as "entered". However, pro se litigants and the public at large, who must rely on PACER access alone, would not be able to discern the difference between records where valid NEFs were issued, and ones where invalid NEFs were issued by the courts.[xxix]

Moreover, nowhere in court dockets of any of the courts was there any statement of certification invoking the authority of the clerk of the court, and in various cases it was found that unauthorized personnel executed transactions in the dockets. For example—some 15 key records[xxx] in the docket of Richard Fine v Sheriff of Los Angeles County (2:09-cv-01914) were found to have been 'entered' by a court employee who was not authorized as a Deputy Clerk. Moreover, the Clerk of the latter court refused to certify the docket of the latter case, or to state that the dockets in the latter case and other similar cases were dockets constructed pursuant to authority of the clerk in compliance with US law.[xxxi] Similar requests, which were forwarded to the Chief Judge of the respective court—to investigate the matter and initiate corrective actions, were never answered at all. The PACER dockets of the US courts of appeals, likewise, never named the individuals who constructed the dockets and their respective authority. Moreover, no records were stated as 'entered' in the dockets of the US Court of Appeals, 9th Circuit.

Numerous Orders of the US courts of appeals in various cases were served and posted in the dockets with no verification by any judge at all. Moreover, such orders of the US courts of appeals were served with no NDAs at all.[xxxii, xxxiii] Therefore, the dockets, orders, and judgments of the US courts of appeals were deemed inherently ambiguous[xxxiv, xxxv].

Discussion

The findings described in this study must raise serious concerns regarding substantial ambiguity introduced into the conduct of the US courts and US courts of appeals through the transition to digital administration of the courts. Such ambiguity, in and of itself, should be deemed as antithetical to due process and fair hearings, and undermining the integrity of the courts. The effects of the changes in the administration of the US courts documented in the current report on the protection of human rights and enforcement of the law cannot be overstated. Gains in

integrity of the justice system, which were achieved through generations of struggle, have been lost over the past decade through the introduction of PACER and CM/ECF.

Published Local Rules of Court

The results showed that neither the US district courts, nor the US courts of appeals, published local rules of court to spell out their new procedures, which were adopted by these courts as part of implementation of PACER and CM/ECF. For example the Clerk of the Court, US District Court, Central District of California, claimed that the new procedures were established in General Order 08-02. However, general orders are not the legitimate vehicle for establishing new rules of court. Moreover, General Order 08-02, upon review by a reasonable person would be found vague and ambiguous in defining the new practices of digital administration of the courts. Furthermore – General Order 08-02 was published by the Court as an unsigned order with no name of its author and his/her authority. The Unofficial Anderson Manual of the same court provides additional details regarding the new procedures of the Court. However a third party "unofficial" publication could never be deemed a legitimate vehicle for Local Rules of Court in compliance with the law.

The US Rulemaking Enabling Act was signed into law in 1934, authorizing the US Supreme Court and to lesser degree other US courts to promulgate the Federal Rules of Civil/Criminal Procedure. The Rulemaking Enabling Act never mandated the courts to publish new rules. However, failure to publish rules pertaining to fundamental court procedures must be seen as contradictory to due process and fair hearings. Rulemaking by the US courts has a long history as a bone of contention. The Federal Rules of Evidence promulgated by the US Supreme Court in 1972 raised substantial objections, which eventually led the US Congress to enact substantial modifications of the rules. Therefore, it appears that when it came to establishment of PACER and CM/ECF where major changes were introduced in court procedures, the courts took the approach of never publishing any rules at all.

Any computer program by definition is an assembly of assertions, or rules. Therefore, the case management systems of the courts inherently embedded in them numerous rules. Regardless of the fact that such rules were written in computer language of some kind, the courts were required to publish them as rules of court in natural language as well. The US courts deliberately kept such rules hidden from the public and from attorneys appearing before the courts.

Public Access to Judicial Records

Effectively, the setup created by the US courts through PACER and CM/ECF segregated litigants into two separate and unequal groups.[xxxvi] Those assigned to PACER access alone were denied access to critical judicial records—the NEFs and NDAs. Therefore, such litigants and the public at large were unable under current conditions to distinguish between authenticated and unauthenticated judicial records—those which were deemed by the courts themselves as valid and effectual,

versus those which were deemed by the courts themselves as void, not voidable. Needless to say, such a setup at a minimum limited PACER users to a vague and ambiguous perspective on judicial records. The evidence (an example is provided from the Middle District of Florida) showed that even competent experienced attorneys failed to understand the new unpublished rules, and showed the inherent inability of such attorneys to distinguish the validity, or lack thereof in particular court records.[xxxvii]

The failure to define the NEFs in Local Rules of Court in the California Central District and other US courts, combined with their vague and ambiguous description in General Order 08-02, their universal exclusion of the NEF from PACER, and the routine denial of access to NEFs by clerks, were unlikely to be viewed by a reasonable person as merely coincidental. In parallel, the US courts and courts of appeals were shown in the current report to routinely engage in posting in PACER dockets, and in the case of the US district courts, also listing as "entered", papers which were either unverified or never authenticated and therefore could never be deemed by the courts themselves as valid and effectual judicial records. Through such concerted actions, each and all of them in contravention of the law, the public at large, and pro se litigants in particular, were misled into assuming that various records which were deemed void by the courts, were in fact the law of the land.

Judicial Records–Verification and Authentication

The current report documents vague and ambiguous conditions regarding the nature of valid and effectual verification by judges and authentication by clerks of judicial records. Regardless, the finding clearly documents the issuance of records that were never verified and/or authenticated, but were presented as such in the PACER records.

The refusal of the Clerk and Chief Judge of the US District Court, Central District of California, to investigate allegations of misconduct at the Court relative to the issuance of invalid, void, unauthenticated judicial records, and their posting in PACER dockets as "entered", or to initiate corrective actions, provided further evidence that the conduct of the courts in this regard did not result from inadvertent errors.

A further review of the NEF and NDA[xxxviii] forms raised grave concerns. In particular courts, e.g. US District Court, South Carolina, a feature was implemented which made the link between the NEF and the respective record expire after 15 days. There could not possibly be an explanation for such a feature that would be consistent with valid authentication practices of court records.[xxxix]

Comparison of standard authentication counterparts, e.g.–Certification of Acknowledgement by notary public,[xl] and an older, paper-based Certificate of Mailing and Notice of Entry by Clerk,[xli] to the NEF and NDA, revealed additional inherent defects with respect to four basic features: a) title, b) certification statement, c) relationship to the certified record, and d) signature/authority.

The two former forms included in their titles the word "Certificate" or "Certification", while the NEF and NDA failed to include such words in their titles. The former forms included the key statement "I certify..." while the NEF and

NDA included no such statement. The former forms, upon execution were directly, physically, attached to the record being authenticated (in some jurisdictions to this date—through ribbons and sealed wax or embossed foil), but the NEF and NDA were entirely detached from the authenticated records, and were instead hyperlinked—a practice which was explicitly prohibited by the US District Court, Central District of California, General Order 08-02. In at least in one case, Shelley v Quality Loan Services (09-56133) at the US Court of Appeals, 9th Circuit, a false hyperlink was allegedly employed to generate an invalid, void NDA.[xlii] The Court refused to correct the defect even upon request by the Appellant. The former forms, when executed, included "wet" graphical signatures in traditional signature boxes, where the name of the individual executing the authentication was typed below the signature line, and his/her authority to certify the authentication record was spelled out. In addition—a stamp or an imprint of a personal seal or a court stamp were affixed. The NEF and NDA fail to name the individual who purportedly executed any authentication, the authority of the clerk of the court is never invoked, and no stamp or seal or any personal signature of any kind appear in the NEF and NDA that could possibly be deemed as indicating an intention to take responsibility. A checksum string, in and of itself, carries no such significance at all. In short—the NEF and NDA, as implemented in CM/ECF, failed to make any claim of certification of authentication of a specific court record, and failed to convey any intention to take responsibility by an individual invoking the authority of the Clerk of the Court.

The Validity, or Lack Thereof, of PACER and CM/ECF as a Whole

Based on the deficiencies described above, a reasonable person is likely to conclude that the NEFs and NADs are inherently vague and ambiguous, if not void, and thus undermine the validity of the CM/ECF court records as a whole. Furthermore, a valid case management system should never have permitted the issuance of court records and inclusion of such records in court dockets while no valid NEFs or NDAs bearing an electronic "Document Stamp" were issued—if the courts considered such a checksum string as being of any significance. Likewise, a valid and secure case management system should never have permitted persons who were not authorized as deputy clerks to access court dockets and execute transactions that should have been permitted only to authorized persons. Combined with other deficiencies it is practically certain that PACER and CM/ECF as a whole would never have been deemed valid case management systems had the systems been subjected to publicly and legally accountable validation—certified, functional logic verification.

The law[xliii] requires US judges to "initiate appropriate action when the judge becomes aware of reliable evidence indicating the likelihood of unprofessional conduct by a judge or lawyer." However, the evidence in the current report documents the refusal of judges to correct false and misleading court records. Moreover, it is unreasonable to assume that none of the judges in the US, who are trained in the law and immersed in administration of the courts, noticed the glaring defects in PACER and CM/NEF.[xliv, xlv] Therefore it is reasonable to conclude that the compromised integrity documented in the current report was intentional. An abun-

dance of cases involving financial institutions among the cases that were compromised would likewise give conditions at the US courts as one of the fundamental causes of the current financial crisis and dysfunctional state of US banking regulation.[xlvi]

With it, the systems as they are present unique possibilities. Had the US courts permitted public access to the NEFs and NDAs, as required by law, it would be possible to construct a machine generated "Index of Judicial Corruption"—for any US judge who sat on the bench in the past decade, based on the issuance of invalid authentication records, with no understanding at all of the legal matters involved.

Certified Logic Verification and Data Mining are Keys to the Solution

The compromised integrity of the US courts is linked in the current study to the transition of the courts to digital administration. However, such an outcome was not inherent in the transition. On the contrary, digital technologies could provide case management systems that would enhance the integrity and transparency of the courts.

The following are proposed as guidelines for corrective measures:

a) Online public access and case management systems of the courts, which are critical for the safeguarding of human rights and the democratic nature of society, must be subjected to publicly and legally accountable validation (certified, functional logic verification)[xlvii] in all stages of implementation and maintenance. Verification in general, and in relational database management systems in particular,[xlviii] is in principle an NP-complete problem. However, the systems in question are not of such a complexity level as would prohibit functional verification. Moreover, functional logic verification of such systems must be of the highest priority. Therefore any unnecessary complexity must be avoided. User defined integrity constraints must be precisely specified and implemented, and all stages of software implementation must be subjected to structured programming approaches,[xlix] to make them readily amenable to verification.

- Professionals who are versed in computer science and also in the basics of the law—in particular—the law as it pertains to court administration - must undertake verification of such systems. Legally and publicly accountable certification of logic verification implies a process, similar to that which is practiced today relative to building plans, which are subject to public scrutiny in various public planning and zoning boards, and later—through scrutiny of the completed civil engineering projects by authorized inspectors. Within the context of such public process, it is assumed that new principles of public logic documentation and representation would evolve over time as 'standards of care'.

b) Online public access and case management systems of the courts must be required to allow a high level of transparency of judicial records and the systems as a whole. Transparency of the systems at present is limited, and varies considerably among the various courts, with no foundation in the law. Inherent in transparency is also the requirement that all Local Rules of Court, which

are implemented in such systems, be explicitly published in natural language, and posted for public comment and challenge, as required by the US Rulemaking Enabling Act. The only examples presented in the current report involved the failure to publish rules, which were embedded in PACER and CM/ECF, and pertained to verification and authentication of records. However, numerous other rules were found embedded in the systems, which were never published. Further transparency should be required, to allow zero-knowledge monitoring of integrity of the systems through data mining by computing/informatics professionals and the public at large.

c) The public at large must be educated that engagement in data mining of systems that are critical for the safeguarding of human rights and the democratic nature of society is a fundamental civic duty. Case management systems of the courts should be required to permit a high level of transparency, both of the public records inherent in them, and also transparency to allow zero-knowledge monitoring through data mining.[1]

Conclusions

The transition to digital administration entailed a sea change in procedures of the US courts. The transition took place over a relatively short time, and was independently executed by the US judiciary, with insufficient public and legal accountability. The transition resulted in a precipitous deterioration in the integrity of the courts, which undermined the safeguarding of human rights and enforcement by regulatory agencies in the United States. The conditions that were generated as a result are unprecedented in democratic societies in the modern era. They are employed for deprivation of the rights of the people, and to benefit those in government and large corporations. The proposed solution should involve publicly accountable validation (certified, functional logic verification) of case management systems of the courts, system transparency, and ongoing data mining—a civic duty and a prerequisite for the integrity of the courts in the digital era. Although the current report documents conditions at the US courts, similar risks are faced by other nations as well. The international computing/informatics community should assume a leading role in the protection of rights and the democratic nature of society in the digital era.

References

[i] The matters covered in the current study were covered in greater detail in a series of three papers, as a solicitation of expert opinions: `http://www.scribd.com/doc/29525744/`, `http://www.scribd.com/doc/29525890/`, `http://www.scribd.com/doc/29527583/`.

[ii] Codd, E F: The relational model for database management: version 2, ACM Classic Books Series, Addison-Wesley Publishing Company, Inc (1990)

[iii] US Constitution and Acts of Congress, pertaining parts, see—NEF—A Review: `http://www.scribd.com/doc/24732941/`

[iv] *Federal Rules of Civil/Criminal Procedures*

[v] *Rulemaking Enabling Act* 28 USC § 2071–2077: `http://inproperinla.com/10-06-18-rulemaking-enabling-act-s.pdf`

[vi] In US law the right to access judicial records is considered a First Amendment right, as reaffirmed in *Nixon v Warner Communications*, Inc (1978), 435 U.S. 589 (1978). In British law it is deemed a Common Law right.

[vii] US Constitution and Acts of Congress, see–NEF a review: http://www.scribd.com/doc/24732941/

[viii] *Federal Rules of Civil/Criminal Procedures*

[ix] F. Coenen: Verification and Validation Issues in Expert and Database Systems: The Expert Systems Perspective, dexa, 16 pp., 9th International Workshop on Database and Expert Systems Applications (DEXA'98), (1998)

[x] PACER login page: https://pacer.login.uscourts.gov/cgi-bin/login.pl?court_id=00idx

[xi] CM/ECF of the US District Court, Central District of California–login page: https://ecf.cacd.uscourts.gov/cgi-bin/login.pl

[xii] The US Constitution Article I, §9, clause 2: http://www.archives.gov/exhibits/charters/constitution.html

[xiii] NEFs–A review, including samples from various courts: http://www.scribd.com/doc/24732941/

[xiv] Authentication of court records serves dual purpose: (1) Authentication by the clerk of the record duly filed by the judge, or by a party, and entered by the clerk into a given court docket, and (2) Certification that due notice was given to all parties of the entry of the new record.

[xv] artial list of courts and cases, where data were derived for the current report: http://inproperinla.com/10-06-23-partial-list-of-courts-and-cases-s.pdf

[xvi] Access to Local Rules of Court, General Orders, and other specific information of the US District Court, Central District of California: http://www.cacd.uscourts.gov/

[xvii] *Local Rules of Court* of the US District Court, Central District of California: http://www.scribd.com/doc/28862438/

[xviii] *Local Rules of Court* of the US Court of Appeals, 9th Circuit March 30, 2010 records, which were downloaded from the web site of the US Court of Appeals, 9th Circuit: the Local Rules of Court, the General Orders, the CM/ECF User's Guide, and Transcripts of CM/ECF instructional videos. http://inproperinla.com/10 03 30 nda at 9th cca rules orders manuals video transcripts s.pdf

[xix] *General Order 08 02* of the US District Court, Central District of California: http://www.scribd.com/doc/27632471/

[xx] *General Orders* of the US District Court, Central District of California: http://www.scribd.com/doc/28861084/

[xxi] Pertinent parts of the *General Order 08-02* of the US Court, Central District of California: http://inproperinla.com/10-06-18-pertinent-parts-of-general-order-08-02-s.pdf

[xxii] General Orders, CM/ECF User's Guide, and Transcripts of CM/ECF instructional videos of the US Court of Appeals, 9th Circuit: http://inproperinla.com/10 03 30 nda at 9th cca rules orders manuals video transcripts s.pdf

[xxiii] US District Court Central District of California CM/ECF "Unofficial Anderson Manual" http://www.scribd.com/doc/28869533/

[xxiv] See xxii, above.

[xxv] Declaration regarding attempt to access court records at the US District Court, Central District of California. http://www.scribd.com/doc/28932321/

[xxvi] *Richard Fine v Sheriff of Los Angeles County* (2:09-cv-01914) at the US District Court in Los Angeles, California–PACER docket: http://www.scribd.com/doc/32567529/

[xxvii] Unauthenticated court orders of the US District Court, Vermont, in *Huminski v Rutland Police Department* (1:99-cv-160): http://www.scribd.com/doc/25112472/

[xxviii] Two NEFs in *Richard Fine v Sheriff of Los Angeles County* (2:09-cv-01914): http://inproperinla.com/10-06-18-two-nefs-in-fine-v-sheriff-2-09-cv-01914-s.pdf

[xxix] The PACER docket texts for the two court papers, whose NEFs were reproduced in xxxviii, above: http://inproperinla.com/10-06-18-docket-text-for-two-records-in-fine-v-sheriff-s.pdf

[xxx] List of 15 records in *Fine v Sheriff* "entered" in the docket by "dt"–a court employee, but not deputy clerk: http://inproperinla.com/10-06-18-list-of-15-records-entered-by-dt-in-fine-v-sherif-s.pdf

[xxxi] Requests filed with Clerk of the Court Terry Nafisi regarding validity of the docket in Fine v Sheriff: http://inproperinla.com/10-01-17-req-responses-by-clerk-of-us-court-terry-nafisi-re-integrity-of-dockets-in-zernik-v-connor-and-fine-v-sheriff-s.pdf

[xxxii] June 30, 2009 *Order* by the US Court of Appeals, 9th Circuit, in the *Petition–Fine v Sheriff* (09-71692): http://inproperinla.com/10-06-18-june-30-2009-order-of-%20us-court-of-appeals-9th-fine-v-sheriff-s.pdf

[xxxiii] Orders in the Appeal–*Fine v Sheriff* (09-56073): See under index at: http://inproperinla.com/

[xxxiv] Pacer docket of *Richard Fine v Sheriff of Los Angeles County* (2:09-cv-01914) at the US District Court in Los Angeles, California: http://www.scribd.com/doc/32567529/

[xxxv] *Richard Fine v Sheriff of Los Angeles County* (09-56073) - appeal at the US Court of Appeals, 9th Circuit–Docket of the appeal http://inproperinla.com/00-00-00-us-app web-ct-9th-fine-v-sherifh-of-la-0-a-10-03-27-docket-of-appeal-09-56073.pdf

[xxxvi] Separate and unequal: http://inproperinla.com/10-06-18-separate-and-unequal-s.pdf

[xxxvii] Correspondence with Attorney Jack Thompson and Attorney Bob Hurt regarding records of the US District Court, Middle District of Florida: http://inproperinla.com/10-06-07-thomson-v-florida-bar-6-10-cv-442-authentication-s.pdf

[xxxviii] Invalid NDA of the US Court of Appeals, 9th Circuit, on the February 18, 2010 *Mandate* on the *Appeal* from *Petition for Writ of Habeas Corpus* at the US District Court, Central District of California. The *Mandate* itself was not verified by the circuit judges. http://inproperinla.com/10-06-18-nda-of-the-us-court-of-appeals-on-mandate-in-fine-s.pdf

[xxxix] User's Manual for CM/ECF from US District Court, South Carolina, describing a feature where the linkage between the NEF and the respective court record expires after 15 days: http://inproperinla.com/00-00-00-us-dist-ct-a-pacer-cm-ecf-a-cm-ecf-manual-us-dist-ct-nc-faq_v2-2007.pdf

[xl] Sample Certificates of Acknowledgement by public notary: http://inproperinla.com/10-06-16-notary-public-certification-of-acknowledgement.pdf

[xli] Text of Certificate of Mailing and Notice of Entry by Clerk: http://inproperinla.com/10-06-18-text-of-certificate-of-mailing-and-notice-of-entry-s.pdf

[xlii] In the case of *Shelley v Quality Loan Services* (09-56133) at the US Court of Appeals, 9th Circuit, the Order denying petition for a stay, was unsigned, and the NDA was invalid–hyperlinked to a false order. Moreover, the court refused to correct the error upon request by Appellant: Order: http://inproperinla.com/00-00-00-us-app-ct-9th-shelley_09-10-22-a-order-denying-petition-for-stay-unsigned.pdf
Email correspondence with Mr Shelly regarding order, where NDA was hyperlinked to a false record, and the US Court of Appeals, 9th Circuit refused to correct a false NDA: http://inproperinla.com/00-00-00-us-app-ct-9th-shelley_09-10-22-dubious-nda-10-03-26-shelley-email-notice-re-dubious-oct-22-2009-order-and-nda.pdf
False NDA, where the NDA was linked to an unrelated order from a prisoner's habeas corpus petition, still listed as the respective record in the NDA: http://inproperinla.com/00-00-00-us-app-ct-9th-shelley_09-10-22-dubious-nda-9th-circuit.pdf

[xliii] *Code of Conduct of US Judges, Canon 1, Canon 3B(3)*

[xliv] Repeat requests for the Clerk of US District Court, Los Angeles, to state that the docket in *Fine v Sheriff* was valid and effectual in compliance with US law were never answered: http://inproperinla.com/10-01-17-req-responses-by-clerk-of-us-court-terry-nafisi-re-integrity-of-dockets-in-zernik-v-connor-and-fine-v-sheriff-s.pdf

[xlv] Shelley v Quality Loan Services (09-56133)—see xlii, above.

[xlvi] See xv, above.

[xlvii] F. Coenen, see ix, above.

[xlviii] Codd, see viii, above.

[xlix] Dahl, O J, Dijkstra, E W, Hoare, C A R: Structured programming, ACM Classic Books Series (1972)

[l] Robling Denning, D. E.: Cryptography and data security, Addison-Wesley Publishing Company, Inc (1982)

Acknoledgement

The author is grateful for helpful discussions with human rights experts and computer science academic faculty.

Data Mining as a Civic Duty—Online Public Prisoners' Registration Systems

Joseph Zernik
Human Rights Alert (NGO), PO Box 526, 91750 La Verne, California, U.S.A.
e-mail: jz12345@earthlink.net

Abstract

Prisoners' registration systems in the United States are government-controlled networks holding public records that are critical for the safeguarding of liberty. The current report investigated validity, verification, and security concerns pertaining to the Los Angeles, California, online Inmate Information Center. Hundreds of entries were sampled and about half were found invalid. In particular cases access to the arrest and booking records—public records by California law—was requested. Access was denied. Neither were invalid records corrected upon request. Therefore it was concluded that invalid records posted online were not the outcome of inadvertent errors. Similar deficiencies were found in the prisoners' registration system of Marin County, California. Solutions are proposed, based on structured programming and certified, functional logic verification, which must be mandated in such systems. Data mining will remain a civic duty—in the US and worldwide—to safeguard human rights in the digital era.

Key words

functional logic verification; relational databases; Case Management Systems; human rights; prisons and prisoners; register of prisoners; Los Angeles; California; United States; justice system; law; fraud; corruption; false imprisonment

Biographical note

Professor Zernik served on the faculty of the University of Connecticut, University of Southern California, and the University of California, Los Angeles.

In 2010 he founded Human Rights Alert (NGO), dedicated to discovering, archiving, and disseminating evidence of human rights violations by the justice systems of the State of California and the United States in Los Angeles, California, and beyond. Special emphasis is given to the unique role of computerized case management systems in the precipitous deterioration of the integrity of the justice system.

Introduction

Long standing traditions in courts and the justice systems originating from Western Europe require careful public record keeping of prisoners held by the authorities, to prevent abuse—which could result in the deprivation of liberty and oppression of opposition to any prevailing regime.[i] The current study employed data

mining to investigate the integrity of the records in the online prisoners' registration systems of Los Angeles and Marine County, California. In both cases the records provided online in the prisoners' registration systems were not amenable to authentication, and a large fraction of the records were found to be apparently invalid.

Solutions are proposed, which are based on structured programming and publicly and legally accountable validation (certified, functional logic verification). With these—transparency of such systems must be required, and data mining will remain critical for the safeguarding of human rights.

Conditions now prevailing in Los Angeles County, California, as documented in the current report based on analysis of the prisoner registration system, are consistent with previous official and unofficial reports based on lengthy legal investigations which documented large-scale false imprisonment in Los Angeles County. The current report documents that data mining provides low-cost, fast, and effective means for monitoring the justice system. The computing and informatics community is called upon to take a leading role in monitoring human rights in the digital era.

Habeas Corpus—Imprisonment Must Conform with the Fundamentals of the Law

The right to petition for a writ of habeas corpus was established in the English Magna Carta (1215)—whereby any prisoner and/or others are permitted to challenge his/her imprisonment by requesting to be brought before a judge to review the legal foundation for the confinement, and seeking a writ for his/her release in its absence. The US Constitution Article I, § 9, clause 2, states:

> *The privilege of the writ of habeas corpus shall not be suspended, unless when in cases of rebellion or invasion the public safety may require it.*

The late US Supreme Court Justice Louis Brandeis (1856–1941) referred to it as the greatest achievement of the English-speaking legal system – establishing liberty by law. The late Justice William Brennan Jr (1906–1997), referred to it as the "Cornerstone of the United States Constitution". In Fay v Noia (1963), he wrote for the majority of the US Supreme Court:

> *The basic principle of the Great Writ of habeas corpus is that, in a civilized society... if the imprisonment cannot be shown to conform with the fundamental requirements of law, the individual is entitled to his immediate release.*

At minimum "conforming with the fundamental requirements of the law" entails basing the confinement on valid and effectual judicial records—valid and effectual booking records establishing admission of any prisoner into custody of the authorities, which refer in turn to a valid warrant for the arrest, referring to a valid pending court case, or conviction/verdict when imprisonment is based on a settled case.

The Universal Declaration of Human Rights—ratified International Law, likewise prohibits arbitrary arrests and imprisonment.[ii]

Figure 1:
Historic, paper-based Register of Prisoners, City of Santa Monica, California

Prisoners' Arrest and Booking Records–California public records by law

Obviously, no meaningful habeas corpus right could be practiced, if no access was permitted to judicial records that form the presumed foundation for the imprisonment. Therefore, one must consider the right of habeas corpus and the right of public access to judicial records and access to an honest register of prisoners–to inspect and to copy, as closely related fundamental human rights–both of medieval origins. In the United States the right to access judicial records is considered a First Amendment right. In the British legal system it is deemed a common law right. State of California law defines the arrest and booking records of all prisoners as Public Records–California Public Records Act, California Government Code § 6254(f)). The California Public Records Act states:

> *… public records are open to inspection at all times during the office hours of the...agency and every person has a right to inspect any public record… [and to receive] an exact copy.*

In the past, conformity with such a legal framework was accomplished through the maintenance by the authorities of constantly updated Registers of Prisoners (Figure 1), and the maintenance of files holding the respective arrest and booking records, and matching of such records with routine counts of prisoners on location.[iii]

Registers of Prisoners in the Digital Era

With the transition to administration of the justice system based on digital records, the Los Angeles County Sheriff's Department established a setup which is routinely found in other parts of the justice system in California and the US: The legal records are internally maintained through a case management system–a subtype of database management systems. External public access is provided through a sep-

Figure 2: The online public access system of Los Angeles County Sheriff's Department–Inmate Information Center (IIC)

arate online public access system which is a derivative projection of the database but does not present any images of the original legal records. Therefore, the two systems can be viewed as relational databases. Such systems can likewise be viewed as government-owned and regulated social networks. The main objective in data mining of such networks is the safeguarding of human rights.

Objective

The current report investigates the prisoners' registration system of the Los Angeles County Sheriff's Department through data-mining to assess the validity of records presented in the system. The report further assesses the compliance of the system of the Los Angeles County Sheriff's Department with the California Public Records Act, and with fundamental human rights.

Additionally, the results from the Los Angeles County Sheriff's Department were compared with similar results from the Sheriff's Department of Marin County, California.

Finally–the report assesses the role of data mining and computing professionals in the monitoring of government-run networks and the protection of human rights in the digital era.

The System

Online Public Prisoners Registration Systems

The Los Angeles County Sheriff's Department established an online, public access system to the prisoners register–the Inmate Information Center (IIC) (Figure 2).[iv]

Internal Case Management System for Booking Prisoners

The booking records themselves are produced and accessible through a networked, high-security system of booking terminals placed at various law enforcement stations in Los Angeles County, California. The terminals are capable of capturing booking data including demographic data, booking photographs, and finger prints, and link them with warrant, conviction, and sentencing records, as well as listing of prison terms and future scheduled court appearances–data which are derived from court records.

Methods

Data Mining

Data were mined through routine, manual public access to IIC, as permitted by law. Additional access was attempted through the VINE (Victim Information and Notification Everyday)[v] system—a national United States system aimed at providing crime victims with access to prisoners' data.

Assessment of Data Validity

Data captured were assessed for validity based on criteria including:

(a) Prisoners' records, provided online through IIC, were examined for the presence of valid verification and/or authentication.

(b) Prisoners' records, provided online through IIC, were examined for other basic indicators of integrity, including, but not limited to:

i. Availability of a Booking Number for each and every named prisoner.
ii. Consistency of the name of inmate listed in the individual record with the name used for the query, or correct listing of aliases.
iii. Existence of reference to a valid warrant from a valid court record.
iv. Availability of conviction/sentencing or court appearance data from a valid court case.
v. Continuous graphical correlation between Booking Numbers and the Date of Booking in sample populations derived from IIC.
vi. Existence of consecutive data for consecutive Booking Numbers.
vii. Existence of data entries for known prisoners when accessed from the alternative VINE portals.

Attempts to Access California Public Records and Correct False Records

In particular instances where credible evidence was available that data provided through IIC was invalid, false and misleading, attempts were made to access the California public records, which were the arrest and booking records of the individual prisoners, to corroborate or refute the online IIC data. Such attempts were initially carried out through direct written requests to the Sheriff's Department following specific directions provided by the Legal Director of a California civil rights organization[vi], pursuant to the California Public Records Act, California Government Code § 6254(f). Upon denial of such attempts, additional attempts were made to access the records through inquiries to the Sheriff by the highest Los Angeles County elected officials.

In particular instances where credible evidence was available that records provided through IIC were false and misleading, written requests were filed with the Sheriff's Department to correct the false records.

Comparison to Marin County, California, Prisoners' Booking Log

The methods applied to the IIC were applied also to the Marin County Sheriff's Department Prisoners' Booking Log, albeit, in the latter system, access was permitted to the complete register of prisoners, and random sampling of the prisoners' records was therefore possible.

Results

Data in the current report were manually mined, and therefore limited in scope—only a few hundred prisoners' records were examined, as detailed below, and only limited results and conclusions could be reached.

Data Mining

Access to prisoners' data through the IIC is limited. Access is provided through input of the prisoner's first and last name only, and no access is provided by booking number, by date of arrest, by date of booking, or by arrest and booking location. Therefore, random data sampling was not practicable.

To circumvent such limitations, prisoners' records were sampled and collected through queries by common first and last names such as "Jose Ramirez",[vii] "Jose Rodriguez", [viii] and "John Smith".[ix] The two former names retrieved over a hundred prisoners' entries each, most of them from arrests and bookings, which took place in recent years (most convicted prisoners are held in other facilities which are not administered the Sheriff's Department). Data retrieved from the individual IIC records were complied in a table form (Table 1). In addition the actual records or excerpts from the records of individual prisoners were attached in order to demonstrate the nature of the records.

Data Validity

In all three surveys a large fraction of the entries were found to be missing any Booking Number—rendering such entries apparently invalid. Furthermore, in a large portion of the cases, reference was made to judicial records from various "Municipal Courts". However, Municipal Courts ceased to exist in Los Angeles County, California around 2001[x]—almost a decade ago, whereas the respective IIC records were very recent. Such data were deemed invalid as well.

Additionally, records where the court reference or case numbers were missing, or case numbers were provided such as #000000, or #9999999, were deemed invalid records.

Records where the name of the prisoner listed in the IIC records was substantially different, or entirely unrelated to the name used in the query, but neither name was marked as an alias, were deemed invalid as well. Combined, the fraction of invalid records approached 50% of all data samples.

Consecutive Booking Numbers

No access was permitted in the IIC to query prisoners' data by booking number. However, such access was indirectly provided through the VINE system,[xi] which

claims to derive its data from the Los Angeles County Sheriff's Department. No records at all were found for a large fraction of the Booking Numbers in queries of consecutive numbers. Likewise, no prisoners' records were identified, even in cases where the prisoners were known to be held by the Sheriff.

Correlation of Booking Numbers and Booking Dates

Attempts to correlate Booking Numbers with booking dates routinely yielded evidence of the parallel use by the Los Angeles Sheriff's Department of two numbering series: a) The "Low Series"–with Booking Numbers in the 1,200,000 to 1,400,000 range, encompassing 10-15% of Booking Numbers in the various surveys, and b) The "High Series"–with Booking Numbers ranging in the 2,000,000–encompassing the vast majority of Booking Numbers (Figure 3). Attempts to define common factors in the Low Series numbers–such as date or location of the arrest and booking, were unsuccessful. Given the networked nature of the system, such results must be viewed as an alarming indicator of lack of system validity and integrity.

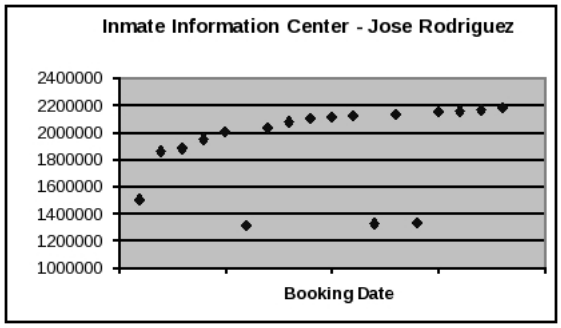


Figure 3:
Correlation of Booking date and Booking Number, results for Prisoner name "Jose Rodriguez"[xii]

Racial Bias, or Lack Thereof

It was assumed that names such as Jose Rodriguez and Jose Ramirez retrieved records that mostly reflected prisoners of Latino ancestry. John Smith was likewise assumed as reflecting records of non-Latino prisoners, albeit that no means was available to distinguish by name "white" and "black" prisoners except by race as listed by the Sheriff's Department. The data accumulated in the current report did not allow the determining of whether any bias existed in the use of false invalid data for any particular ethnic group.

Access to California Public Records

Attempts to access public records, which are the arrest and booking records of Los Angeles County prisoners, pursuant to California law, by direct written requests to the Sheriff's Department, were denied without exception, in disregard of the law. Even when attempts were made to access such records through inquiries by Los Angeles' highest elected officials,[xiii] only false and deliberately misleading records

were produced. Such efforts were focused on cases where the online IIC records were clearly false and misleading.

For example, media and witnesses clearly documented that the 70 year-old former US prosecutor Richard Fine was apprehended on March 4, 2009, at the Superior Court of California, County of Los Angeles, City of Los Angeles, Mosk Courthouse, 111 North Hill Street.[xiv] He has been held in a hospital ward under solitary confinement ever since. In contrast, the online IIC records falsely stated that the arrest and booking took place on the same date—March 4, 2009—but at a location and pursuant to the authority of the non-existent San Pedro Municipal Court, Los Angeles County, City of San Pedro. Moreover, the Sheriff's Deputy at the only court today existing at that location—the San Pedro Annex, Superior Court of California, County of Los Angeles—denied that Richard Fine or anybody else was arrested or booked there in recent years, or that any booking facilities existed at the location at all.

Correction of False IIC Records

The Los Angeles Sheriff's Department was required by law to keep valid and effectual records as the basis for any imprisonment.[xv ,xvi] The Los Angeles Sheriff's Department was repeatedly informed of the false and misleading records posted the IIC system regarding the arrest and booking of Richard Fine and others, and requests were made to correct the data. Regardless, the Sheriff's Department repeatedly produced the false records—unverified and unauthenticated printouts from the IIC, instead of producing a valid arrest warrant and booking records for Richard Fine and others, which were requested. Therefore, it is claimed that a reasonable person would conclude that the posting of false data in IIC in such cases was not the outcome of inadvertent errors. Instead it was a case of fraud by the justice system, intended to affect the false imprisonment of Richard Fine and others.

Comparison to the Marin County Prisoners' Booking Log[xvii]

The Marin County online Booking Log was subjected to data mining similar to that performed in the Los Angeles system, albeit, the system permitted access to all prisoners' records, and therefore random sampling was feasible. The data were found to be far from meeting basic standards of integrity. Over half the records in a sample lacked any reference to court records at all. Moreover, reference was made in such cases to "Confidential Court Cases". No "confidential court cases" are permitted by US or international law. None of the cases lacking reference to court records were those of minors.

Furthermore, no correlation at all was found between Jail IDs and Original Booking Dates of the prisoners (Figure 4).

Upon review of the court cases, in cases which included reference data, all were found belonging to existing cases of the existing Marin County Superior Court.

Table 1: Sample data extracted from the Marin County online Booking Log

#	Page	Name	Jail ID	Original Booking Date	Court Case(s), DOB, Charges	Court Case / Date Filed
1.	6	BALFE, PETER MARSHALL	P00147677	5/9/2010	SC170058A	Found 5/10/2010
2.	7	BARRUS, MICHAEL RAY	P00173530	6/9/2010	No Records for BARRUS, MICHAEL RAY found at this time. (DOB 12/9/1976)	N/A
3.	10	BOISSIERE, DANNY LESHAWN	P00126902	9/16/2009	SC166467A	Found 9/18/2009

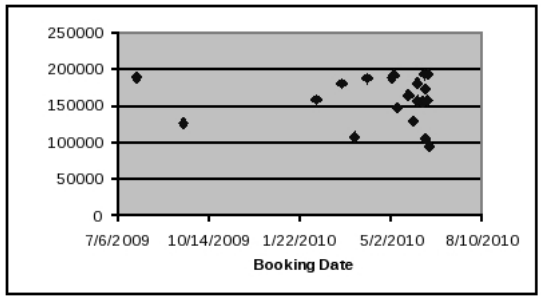


Figure 4:
Jail ID's correlation with Original Booking Dates in Marin County

Discussion

Data Referring to Non-Existent Municipal Courts

The routine use of non-existent court names as the authority for the arrests and booking of Los Angeles County prisoners is of particular concern, since the Sheriff's Department refused to correct such data, even when it was pointed out to the Sheriff's Department that the data were apparently false on their faces. It is hypothesized that the booking terminals and case management system of the Los Angeles Sheriff's Department were established prior to the termination around 2001 of the Municipal Courts in Los Angeles County, California. Therefore, the most plausible explanation for the false "Municipal Court" records is that menus created prior to the termination of such courts were left intact and never updated, and staff routinely select such invalid options, including, but not limited to cases where no valid, honest, and effectual judicial records exist to support the arrest and booking.

Moreover, booking terminals are identified by their location. Therefore, the most plausible explanation for instances such as the false online records pertaining to the former US prosecutor Richard Fine—stating that he was arrested and booked at the non-existent "San Pedro Municipal Court", is that the Sheriff's Department maintains at undisclosed locations terminals which were previously stationed in the municipal courts, even after the respective courts were terminated. Such in-

valid, false and deliberately misleading booking terminals are being utilized to this date—to affect false arrests and false imprisonment—when no legal foundation exists for the confinement.

Assessment of the Scope of False Imprisonment in Los Angeles County, Based on Current Surveys

Los Angeles County, California, is the most populous county in the United States, with over 10 million residents. Accordingly, its county's court and sheriff's department are the largest in the US. Although no direct data were available for the total number of prisoners held by the Los Angeles Sheriff's Department, it could be safely estimated that tens of thousands of prisoners are falsely held in Los Angeles County alone, resulting from arrests in recent years—with no valid and effectual records, as shown in the current report.

Official and Unofficial Reports of Large-Scale False Imprisonment in Los Angeles County, California, and elsewhere in the US

The results presented in the current report lead to the conclusion that the Sheriff's Department of Los Angeles County is engaged in long-term, large-scale false imprisonment of Los Angeles County residents. Such results are consistent with previous official reports regarding the justice system of Los Angeles County, California. As part of the investigation of the Rampart Corruption Scandal (1998–2000), the framing and false imprisonment of many thousands of people, almost exclusively blacks and Latinos, was established.[xviii, xix] Later, a three-year official report (2003–2006) by the Blue Ribbon Review Panel, published in 2006, concluded, "innocent people remain in prison".[xx] Then Dean of Loyola Law School, Los Angeles, David Burcham wrote, "...judges tried and sentenced a staggering number of people for crimes they did not commit."[xxi] The independent report of Prof Erwin Chemerinsky, renowned constitutional scholar and Dean of the University of California Law School concluded, "this is conduct associated with the most repressive dictators and police states... and judges must share responsibility when innocent people are convicted."[xxii] Hardly any of those who were documented as falsely convicted and falsely imprisoned have been released to date. Judges were documented as objecting to their release.

The results presented here are not unique for Los Angeles County, California, either. The data from Marin County raise the same concerns, although no false court names were employed, and a much smaller number of prisoners is involved.

In the ongoing Kids for Cash scandal, currently unraveling in Luzerne County, Pennsylvania, it was exposed that judges collected kickbacks amounting to millions of dollars from privatized jails, in exchange for the false imprisonment of juveniles by the thousand.[xxiii] It is not clear to what degree digital case management and online public access systems facilitated the false imprisonments in Pennsylvania.

It is likely that the justice system and prisons in other parts of the world where transition has been made to digital records are susceptible to abuse as well, similar to that which was documented here in Los Angeles and Marin County, California, in the current report.

Data Mining as the Key to Public Monitoring of the Justice System

Regardless of the limitations in data mining by the Los Angeles County ICC, and the Marin County Prisoners Booking Log, public access provides a powerful tool for monitoring of the justice system. The simple manual surveys conducted in the current report allowed the demonstration of inequities, which are related to large-scale deprivation of liberty and abuse of human rights.

One should notice that even in a prominent case—like the apparent false imprisonment of Richard Fine—a former US prosecutor—mainstream media to this date have failed to base their reporting on direct examination of the integrity, or lack thereof, of the digital records, relying instead on oral pronouncements by various officials. Such circumstances make patent the need for computing specialists to assume a more prominent role in monitoring the justice system.[xxiv] It is proposed that computing professionals have unique skills that would allow them to be in the forefront of human rights protection in the digital era.

Proposed Corrective Measures

The following are proposed as guidelines for corrective measures:

(a) Online public access and case management systems that are critical for the safeguarding of human rights must be subjected to publicly and legally accountable validation (certified, functional logic verification) in all stages of development, through implementation, to any updates and modifications.
(b) Such systems must be required by law to allow a high level of transparency that would allow ongoing effective public monitoring through data mining, as well as a widely distributed, zero-knowledge approach to system security.
(c) The public at large must be educated to assume the data mining of systems that are critical for human rights and the stability of democratic government as a civic duty.

Conclusions

With the transition to administration of the justice system through digital records, the Los Angeles County Sheriff Department established a setup which is routinely found in other parts of the justice system in California and elsewhere in the United States: Legal records are internally held by the authorities in case management systems—a subtype of database management systems, where public access is denied. Public access is routinely provided through a separate, online, public access system. The validity and integrity, or lack thereof, of such a setup of relational databases, is the essence of the deficiencies identified in the current report. The records, which are provided in the online public access system, are neither verified, nor authenticated in any manner at all. Yet the authorities rely on the public's confidence in such records. To compound the problem, this setup of relational databases is employed by the authorities to deny public access to what are public records by law—the honest, true, and valid arrest and booking records of the prisoners.

The results presented in the current report lead to the conclusion that the presentation of such false data in the Los Angeles County IIC was not the outcome of

inadvertent errors, but part of conduct intended to affect false imprisonments and deprivation of Liberty. Beyond the abuse of those who are falsely imprisoned, the mere existence of such conditions in the justice system in Los Angeles and Marin Counties, California, are alleged to be large-scale abuse of the human rights of over 10 million residents of these counties, by the justice system itself.

The findings of the current report are consistent with previous media and official reports of large-scale false imprisonment in Los Angeles County, California, mostly of black and Latino prisoners. The novelty in the current report is only in demonstrating that data mining of such systems allows the public to document the abuses without resorting to complicated and protracted public investigations. Therefore, while digital systems provided simple tools for the justice system to circumvent the law, data mining of the same systems provides a simple and effective tool to demonstrate the corruption of the justice system and large-scale human rights abuses.

Ways and means are readily available whereby the systems fraudulently erected by the justice system authorities with no public oversight at all, could be remedied—through publicly and legally accountable validation (certified, functional logic verification). Regardless, transparency should be required, which would permit routine data mining, which must be viewed as a civic duty. The public at large must perform its duties and obligations and constantly monitor the justice system—to safeguard the integrity of the prisons and protect human rights in the digital era. The computing and informatics community should lead the way.[xxv]

References

[i] Requests for comments, corrections of a draft of this letter were forwarded by fax on June 3, 2010 to the office Lee Baca - Sheriff of Los Angeles County, Charles McCoy—Presiding Judge, and John A Clarke—Clerk of the Superior Court of California, County of Los Angeles. No comments or corrections were received.

[ii] The Universal Declaration of Human Rights, Article 9, states: No one shall be subjected to arbitrary arrest, detention or exile.

[iii] California Code of Regulations, Title 15—Crime Prevention and Correction, Title 15, Article 2, § 3273-4, Article 4, § 1041: `http://inproperinla.com/10-06-13-warrants-booking-penal-code-regulations-s.pdf`

[iv] Los Angeles Sheriff's Department Inmate Information Center: `http://app4.lasd.org/iic/ajis_search.cfm`

[v] VINELink 2.0: `https://www.vinelink.com/vinelink/initMap.do`

[vi] The author is grateful to the unnamed Legal Director for his help.

[vii] Jose Martinez data survey: `http://www.scribd.com/doc/24809956/`

[viii] Jose Rodriguez data survey: `http://www.scribd.com/doc/25064776/`

[ix] John Smith data survey: `http://www.scribd.com/doc/24816245/`

[x] January 2000 Los Angeles Times report of the termination of the Municipal Courts in Los Angeles County, California: `http://www.scribd.com/doc/32446226/`

[xi] Survey of Los Angeles prisoners data through the VINE system: `http://www.scribd.com/doc/25315610/`, `http://www.scribd.com/doc/28350775/`

[xii] See v, above.

[xiii] Correspondence between Los Angeles County Supervisor Michael Antonovich and office of Los Angeles County Sheriff Lee Baca–in attempt to gain access to the arrest and booking records of former US prosecutor Richard Fine–held by the Sheriff in a hospital ward under solitary confinement for the past 14 months, and attorney Ronald Gottschalk - held at the time in psychiatric hospital ward: http://www.scribd.com/doc/25555341/

[xiv] Full Disclosure Network: Attorney Jailed In Attempt to Disqualify L.A. Judge For Taking Bribes, March 4, 2009: http://www.scribd.com/doc/32458545/

[xv] The requirement for an arrest warrant is embedded in the Fourth Amendment to the US Constitution: The right of the people to be secure in their persons, houses, papers, and effects, against unreasonable searches and seizures, shall not be violated, and no Warrants shall issue, but upon probable cause, supported by Oath or affirmation, and particularly describing the place to be searched, and the persons or things to be seized.

[xvi] *California Code of Regulations, Crime Prevention and Correction* 3273 (Acceptance and Surrender of Custody) says: Wardens and superintendents must not accept or surrender custody of any prisoner under any circumstances, except by valid court order or other due process of law.

[vii] Survey of the Marine County, California Prisoners' Log: http://www.scribd.com/doc/32981371/

[xviii] Rampart-FIPs (Falsely Imprisoned Persons): Rampart First Trial, PBS Frontline, Rampart False Imprisonments: http://www.scribd.com/doc/24901612/

[xix] Rampart-FIPs (Falsely Imprisoned Persons). Review: http://www.scribd.com/doc/24729660/

[xx] Blue Ribbon Review Panel: Rampart Reconsidered, The Search for Real Reform Seven Years Later (2006): http://www.scribd.com/doc/24902306/

[xxi] Burcham, D. and Fisk, K.: The Rampart Scandal: Policing The Criminal Justice System, Loyola Law Review, Vol. 34, pp. 537-543 (2001): http://www.scribd.com/doc/29043589/

[xxii] Chemerinsky, E: The Criminal Justice System of Los Angeles County, Guild Practitioner, Vol. 57, pp. 121-133 NLG (2001): http://www.scribd.com/doc/27433920/

[xxiii] Debra Cassens Weiss: Ex-Judge to Plead Guilty in Kids-for-Cash Scandal; Is He 'Singing Like a Bird'? American Bar Association Journal, April 30, 2010, http://www.scribd.com/doc/30840581/

[xxiv] Abbie Boudreau, Emily Probst and Dana Rosenblatt: Ex-lawyer jailed 14 months, but not charged with a crime, May 24, 2010 CNN: http://www.cnn.com/2010/CRIME/05/24/jailed.lawyer.richard.fine/index.html See also xv, above. Both reports relied on interviews and failed to examine or report on the records in the case.

[xxv] Online petition: Free Richard Fine: http://www.thepetitionsite.com/1/free-fine

Short communications

A Proposal for an Approach to Extracting Conceptual Descriptions of Hyper-linked Text Documents

Gabriel Lukáč
Department of Cybernetics and Artificial Intelligence, FEI, TU Košice, Letná 9, 042 00 Košice, Slovakia
e-mail: Gabriel.Lukac@tuke.sk

Abstract

This paper is a brief description of work that is planned to be realised during the PhD study of the author. It describes a proposal for a methodology for extracting conceptual descriptions of text documents published within a social network of authors. Additionally, the hyper-linked nature of particular documents (e.g. a citation network of blogs) provides a very good framework in which to take an advantage from the networked environment and it allows one to use algorithms for information diffusion to track influential concepts spreading down the network. The paper is divided into several sections. It gives an overview of related work that serves as a foundation for the presented method. At the end an outline of a particular method proposal is provided. Its main purpose is to foster further discussions about the topic and serve as a baseline for future work.

Key words

social networks; textual documents; diffusion of innovation, concepts, keyword extraction

Biographical note

Gabriel Lukáč is a PhD student in the Department of Cybernetics and Artificial Intelligence at the Faculty of Electrical Engineering and Informatics, Technical University of Košice. He is interested in research into the dynamics of complex networks with a focus on the spreading of information through networked structures. In 2007 he graduated (MSc degree) at the same faculty.

Introduction

The process of generating conceptual descriptions of textual documents automatically is an important challenge in the semantic web initiative (W3C Consortium, 2010). Without the possibility to make such automatic extractions, the vision of the semantic web will still remain a vision and will never be deployed in practice. There is another important use of such techniques: Keyword extraction for search engine optimisation (SEO). Many SEO companies and practitioners use it as a daily tool to extract keywords. In general, this process is mostly realised for documents, for which their hyper-linked structure is not considered. On the web there is an invaluable source of large-scale document data generated by the social network of authors and data itself has a very clean network structure. We believe that

this advantage of networked structure can provide an opportunity to use principles of diffusion of information (Rogers, 2003) to track important concepts spreading through the network. Additionally this framework can give us the possibility to come-up with a measure that will produce a ranking of the concept importance for a given document.

Related work

In this section work related to our paper is provided. It is divided into two subsections, and each of them represents a topic that will be necessary to revisit in future research.

Spreading of information cascades

Models for catching the spreading of information through information channels are very often based on epidemic models (Satorras and Vespignani, 2001) describing the spread of viral diseases through social networks of people. An analogy to such epidemic models is the spread of so-called *information cascades*. Information cascades are phenomena in which an action or idea becomes widely adopted due to the influence of others, typically neighbors in some network (Leskovec et al., 2007). Every information cascade (or sometimes in literature it is denoted as a *conversation tree*) has one starting node called cascade initiator. In the case of emails or discussion forum posts the *cascade initiator* is the contribution starting an email or discussion thread. In the case of blogs, the cascade initiator is a blog article that comes-up with a certain unique idea for the first time. If the article is interesting enough, it will start an information cascade that will be successively built up by the progressive adding of new articles to the cascade – articles that make cite former blog contributions.

The idea of cascades spreading through networks has been studied from the point of view of various branches of science. Rogers (2003) has studied them as a sociological phenomenon called *diffusion of innovation*. More suitable for the purpose of this paper is the work of Kumar et al. (2004), where it was used to explain actual trends in the blogosphere.

To model the process of adoption of some idea that spreads through an information cascade, two groups of models are usually used: *Threshold models* (Granovetter, 1978), where adoption of an idea by some node (actor) is conditioned by the overall sum of weights of incident edges above a certain threshold t. The second class of cascade models are *independent cascade* models (Goldenberg et al., 2001), in which the chance that node i will adopt the behaviour of node j is given by probability p_{ij}.

Identification of spreading concepts

A critical problem when analysing information cascades is the identification of words, concepts, phrases (or sometimes the concept of a *meme* is used) that are being spread. It is a non-trivial assignment with a great complexity. In the literature there are a number of approaches to solving this problem. In this chapter we will

show one approach that is directly related to networks and two others, that have their foundations in the field of information retrieval and the social web.

The first methodology we want to mention here is the *MemeTracker* of authors Leskovec et al. (2009) for extracting short textual phrases. It is an algorithm creating clusters by aggregating similar textual phrases and it declares this cluster as a standalone cascade. Particular phrases are perceived as an analogy of the 'genetic code' of various memes. This means that the similarity of several memes remains with increasing time still recognisable, but as in genetic structures, so also memes are subject to considerable mutations.

The work of Matsuo and Ishizuka (2003) is also related to the topic of identification of spreading concepts even if it is not related directly through network terminology as a common denominator. Their work has a basis in the theory of *information retrieval*. They based their algorithm for keyword extraction on the idea of frequent co-occurrence of some term with a certain subset of other terms. If one can find such a term in a text, then it is probably a key term of a given text.

Another important (and nowadays very popular) approach for obtaining conceptual descriptions of documents are *folksonomies* – online services for shared categorisation of web resources (Al-Khalifa and Davis, 2006). Their biggest advantage is the relevance of concepts that were contributed by respective users as annotations of their bookmarks. Moreover they provide public API that enables its users to access and download keyword descriptions for every URL that has been marked as public.

Concepts extraction method proposal

The purpose of this method is to provide a list of concepts (keywords) for any given document from the data set. That gives us a very clean list of inputs and outputs of the future algorithm. At the input there is a database of text documents along with an adjacency matrix that will cover citation bindings between particular texts. At the output one should have an opportunity to make a query for any document and obtain a list of the most representative keywords that characterise the document as accurately as it is possible.

We plan to divide our method into two parts. The first one will be the cascades reconstruction part, in which we will try to reconstruct as many cascades from the data as possible. This phase will be based on research on threshold and independent cascade models, trying to fit the data and extract conversation trees. For the second part we will use the MemeTracker methodology (Leskovec et al., 2009) combined with public APIs of online bookmarking services to identify concepts that spread through conversation trees.

Having a list of concepts available for every text document together with the nesting depth saying how deep the concept has spread within the conversation tree, it will be feasible to design a measure that will rank the representativeness of the concept to a particular document.

Conclusion

In this paper we give an overview of the future work that will be realised within the next year of the author's PhD study. Unfortunately it does not give any recent results, but its aim is to provide the foundation to raise discussions about the topic of generating representative concepts of texts and hopefully it will foster new ideas.

References

Al-Khalifa, H. S., Davis, H. C. (2006): Folksonomies versus automatic keyword extraction: An empirical study [on-line]. [cit. 2010-08-05]. Available at: `http://www.iadis.net/dl/Search_list_open.asp?code=2728`.

Goldenberg, J., Libai, B., Muller, E. (2001): Talk of the Network: A Complex Systems Look at the Underlying Process of Word-of-Mouth. *Marketing Letters*, *, pp. 211–223. (ISSN 0923-0645.)

Granovetter, M. (1978): Threshold Models of Collective Behavior. *American Journal of Sociology*, 83.6, pp. 1420.

Kumar, R. et al. (2004): Structure and evolution of blogspace. *Commun.ACM*, Vol. 47, No. 12, pp. 35–39. (ISSN 0001-0782.)

Leskovec, J., Backstrom, L., Kleinberg, J. (2009): Meme-tracking and the dynamics of the news cycle. In: *KDD '09: Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*. Paris (France) : ACM, pp. 497–506. (ISBN 978-1-60558-495-9.)

Leskovec, J., Faloutsos, Ch. (2007): Scalable modeling of real graphs using Kronecker multiplication. In: *ICML '07: Proceedings of the 24th international conference on Machine learning*. Corvalis (Oregon) : ACM, pp. 497–504. (ISBN 978-1-59593-793-3.)

Matsuo, Y., Ishizuka, M. (2003): Keyword Extraction from a Single Document using Word Cooccurrence Statistical Information. In: *Proceedings of the 16th International FLAIRS Conference*. St. Augustine (Florida) : AAAI Press, pp. 392–396. (ISBN 1-57735-177-0.)

Rogers, Everett M. (2003): *Diffusion of Innovations*. 5th ed. New York : Free Press, . ISBN 0-7432-2209-1.

Satorras, R. P., Vespignani, A. (2001): Epidemic spreading in scale-free networks. *Physical Review Letters*, vol. 86, no. 14, pp. 3200–3203.

W3C Consortium (2010): Semantic Web Activity [on-line]. [cit. 2010-08-05]. Available at: `http://www.w3.org/2001/sw/`.

The work presented in this paper was supported by the Slovak Grant Agency of the Ministry of Education and Academy of Sciences of the Slovak Republic within the 1/0042/10 project 'Methods for identification, annotation, search, access and composition of services using semantic metadata in support of selected process types'.

Persuading Collaboration: Analysing Persuasion in Online Collaboration Projects

Ronan McHugh
Royal School of Library and Information Science, Birketinget 6, 2300 Copenhagen S., Denmark
e-mail: mchugh.r@gmail.com

Birger Larsen
Royal School of Library and Information Science, Birketinget 6, 2300 Copenhagen S., Denmark
e-mail: blar@db.dk

Abstract

In this paper we propose that online collaborative production sites can be fruitfully analysed in terms of the general theoretical framework of Persuasive Design. OpenStreetMap and The Pirate Bay are used as examples of collaborative production sites. Results of a quantitative analysis of persuasion in these sites are presented and discussed. This framework may be of value to other researchers interested in design of persuasive systems.

Key words persuasion; online collaboration

Introduction

One of the most striking features about the growth of the Web over the past few years has been the remarkable success of web based services that derive their value from 'crowd sourced' production, that is sites where the majority of content is created by users themselves rather than the companies, individuals or institutions behind the site. But how do these sites convince, or persuade, their users to take part and to remain active, and thus continue contributing content? The focus of this paper is on using a persuasive framework to analyse quantitative data on user behaviour in collaborative sites.

The rise of collaboration

The growth of volunteer based web-based projects producing goods in a collaborative manner has attracted considerable interest within academia. Quantitative analyses of participation have been applied to several different projects including Wikipedia, Flickr and Usenet newsgroups (Nielsen, 2009; Ortega, 2009; Shirky, 2008). These analyses have tended to focus on system-level analyses such as the rate of participation inequality in collaborative projects. In this analysis, we try to tie quantitative data to the level of the individual participant.

Persuasive Design

Recent years have also seen the birth of the field of Persuasive Design, which is concerned with the ways in which computers and related devices can alter user

behaviour through psychological processes. B. J. Fogg (2002), the founder of the discipline, defines persuasion as *an attempt to change attitudes or behaviours or both*. The ubiquity of computing devices makes computer-based persuasion a crucial topic of study for research on digital society. In the present paper, it is argued that the perspective of Persuasive Design can be fruitfully applied to the study of online collaborative projects.

Persuading Collaboration

Collaborative production sites are defined by the fact that the bulk of content is provided by users themselves in collaboration with other users. The role of site developers is therefore to create a platform that allows this creation to take place, as well as the communication and coordination that is involved in collaboration. This task clearly involves persuasive design, as the site design must encourage users to take part and to remain active. Indeed, the persuasive content of such sites may tend to be many times more complex than other examples of persuasive design, as they involve a significantly more complex array of actions including coordination with other users.

Methodology

This study applies a persuasive lens to a quantitative analysis of participation in online collaboration projects. The quantitative analysis is based on complete user histories downloaded from two such projects; Open Street Map (OSM) and The Pirate Bay (TPB). Open Street Map is a collaboratively produced map of the world. Participants contribute by adding points to the map which they may have derived from exploring an area with a GPS transmitter or simply from local knowledge. The Pirate Bay is a site which indexes torrent files which are used to download files collaboratively, from multiple computers at a time. Participants contribute by uploading torrent files and allowing other users to download files from their computer.

Data retrieval

The data for this study was retrieved by downloading histories of user activities stored publicly on the websites in question. URLs for user profiles were obtained by entering the unique subdirectories for user profiles into Yahoo! SiteExplorer (available at `http://siteexplorer.search.yahoo.com/`) and downloading the first 1,000 results, the maximum allowed by SiteExplorer. Duplicates were removed and a script was used to download the full histories associated with each user, converting pages from a html format into a tabbed text file.

Analysing persuasion

In order to analyse persuasion in our cases we need to specify some persuasive goals for such sites. For this study we have specified two goals: (1) encouraging users to participate multiple times and (2) encouraging users to remain active over time.

In order to analyse (1) we constructed frequency distributions in which users were sorted according to the number of times they had participated in; their respective project. A hierarchy was created showing the proportion of users from the sample who had contributed once, twice, three times etc. This hierarchy thus shows the proportion of users who only contribute at these small scales.

In order to analyse (2) we constructed frequency distributions based on total user lifetime, i.e. the number of days between their first and last participation event. These hierarchies thus show the projects' success at retaining users for longer periods of time.

It must be stressed that in comparing participation patterns between systems we are not necessarily comparing like with like; creating a torrent file may involve considerably more work than adding a point to a map although if data is gathered through GPS tracing, this may also involve a considerable amount of effort. For this reason it is important to use a variety of different measures in order to analyse persuasive success or failure.

Analysis and Discussion

The TPB dataset consisted of 268,141 torrents produced by 1,495 users (TPB store usernames at two different locations, which made it possible to download 2,000 urls, of which 1,495 were unique). The set had an average contribution of 179.36 torrents per user with a median of 10. The OSM dataset consisted of 1,884,104 edits contributed by 762 users. This gives an average of 2472.58 edits per user, with a median of 299.

Table 1 compares the proportion of users from each sample contributing at different scales. What is clear from this is the large difference in participation patterns between OSM users and TPB users. An extremely large proportion of TPB users only ever contribute one torrent to the project, while only a small proportion of OSM users do the same. This indicates a persuasive failure on the behalf of the Pirate Bay when it comes to encouraging repeat contributions. Of course, we must remember that contributing a torrent will frequently involve more work than adding a point to a map. For this reason it is worth looking at lifespans of users as another measure of persuasive success.

Table 1: Proportion of users contributing between 1 and 5 times

Number of contributions	OpenStreetMap	The Pirate Bay
1	1.31% (10)	17.24% (258)
2	1.31% (10)	8.35% (125)
3	0.13% (1)	6.41% (96)
4	0.52% (4)	5.08% (76)
5	0.26% (2)	3.00% (45)
≤5	3.54% (27)	40.10% (600)

The average lifetime of TPB users is 308.35 days and the median is 169 days compared to 514.88 days and 516 days for OSM users.

Table 2 shows the proportion of users from each project whose lifetimes last 1–5

Table 2: Proportion of users with lifespan of 1–5 days

Lifetime (days)	OpenStreetMap	The Pirate Bay
1	9.05% (68)	21.67% (324)
2	1.86% (14)	2.47% (37)
3	0.26% (2)	0.86% (13)
4	0.26% (2)	0.66% (10)
5	0.26% (2)	0.93% (14)
≤5	11.71%	26.62% (398)

days. As is apparent, OSM editors tend to remain involved with the project longer than TPB users. This indicates that the OSM project is better at persuading users to remain active than TPB is.

Conclusions and future work

This paper examines some ways in which quantitative data can be tied to a persuasive analysis of collaborative projects. It suggested two measures of persuasive success, user contributions and user lifetime. Many other analyses are possible and ideally these should be tied to a heuristic analysis of site features.

References

Fogg, B. J. (2002): *Persuasive Technology: Using Computers to Change What We Think and Do.* San Francisco : Science and Technology Books, 283 pp.

Nielsen, J. (2006): Participation Inequality: Encouraging more users to contribute [on-line]. [cit. 2010-06-05]. Available at: `http://www.useit.com/alertbox/participation_inequality.html`.

Ortega, J. (2009): *Wikipedia: A Quantitative Analysis.* (PhD thesis.) Madrid : Universidad Rey Juan Carlos. 228 pp.

Shirky, C. (2008): *Here Comes Everybody: How change happens when people come together.* London : Penguin Books, 344 pp.

Methodics of Using WEB 2.0 Services for Higher Education

O. M Samolyenko
*
Ilona V. Batsurovs'ka
South street, house 54, ap. 11 (private address), 54034 Nikolaev City, Ukraine
e-mail: ilona82@inbox.ru
N. S. Ruchins'ka

Abstract

The article examined services WEB 2.0 and methodics of their using in the educational process of higher school.

Key words

WEB 2.0; methodic of using WEB 2.0 services; blog; communication; filling of content; higher education

Statement of the problem

National doctrine of development of education points to priority of inculcation of contemporary informational-communicative technologies that provide further perfection of educational and upbringing process and preparation of young generation to life activity in the informational society. Development of individuality must be done in the realities and in the perspectives of informational dynamics, habits of self-made scientific cognition, self-education and self-realisation (Decree. . . , 2002).

Higher education must provide rise of individuality of a student by such kind of methods and forms of learning that would develop his/her active attitude towards future professional activity, to generate his/her desire to self-perfection and self-development.

Contemporary informational society demands from future experts mastering of great quantity of information. But the main problem is in the insufficient regulation and systematisation of information concerning arisen needs.

It is possible to solve this problem by way of using services WEB 2.0, that will provide access to regulated and systematised information, and will give opportunity for optimum organisation of learning and self-dependent work.

Analysis of the last researches and publications: Questions connected with self-dependent work in the sphere of informational process and with creation of Web-oriented environment were investigated by such scientists as V. V. Olyinik, L. L. Lyahots'ka, V. M. Kuharenko, O. V. Rybalko, N. G. Syrotenko. Questions devoted to separately self-dependent works were examined by P. I. Pidkasistyi, T. K. Kuchera, T. M. Kartel and others.

Formulation of the aim

The aim is to examine notions of services WEB 2.0 and possibilities of their using in pedagogic activity and to describe methodics of using blogs for support of educational process in higher education.

Statement of the main material

Forming of virtual educational spaciousness was provoked by the desire to connect present pedagogical experience with new informational technologies. Using of services WEB 2.0 for higher educational plays a key role in the process of such kind of a connection. Key factors of perfection of technologies WEB 2.0 is an open character of informational filling, speed of access and their placing, independence from individual schedule of involvement of participants into the process of communications during the time of joint work.

Services WEB 2.0 open new possibilities for activity both for lecturers and students: from the search of the information on the internet to creation and editing of their own digital objects-texts, schedules, programs, audio and video records and others.

Peculiarity of WEB 2.0 is the principle of involvement of users for filling and non-expendable measuring of content.

The appearance of the title WEB 2.0 we can connect with the article of Tim O'Reilly 'What is Web 2.0' (2005); firstly it was published in Russian language in the journal 'Computerra' (2005) and than it was placed on internet under the title 'What is Web 2.0' by web-site of 'Computerra online'. In this article O'Reilly (2005) connected the appearance of big quantity of sites, united by some general principles with general tendency of the development of internet-associations, and he named this event WEB 2.0 to distinguish it from the out of date WEB 1.0.

For the main principles of WEB 2.0 we can relate to the inalienable right of users to create content self-dependently, to manipulate it and to manage the ties between their own and foreign materials; thus we can speak about a coordinated activity of separate users that form and fill the net with their content. Such activity is characterised by the raised level of communication, coordination and involvement of users in the process of using and creation of resources, replenishment of services, and determination of strategy of development of resource in general.

For the main pedagogical opportunities of using services WEB 2.0 for higher education we can relate to:

- Using of open, free of charge and free electronic resources by students and lecturers. During the process of broadening of social services we can see the accumulation of materials that can be used with learning aim;
- Self-dependent creation of net learning content. Lecturers and students have the possibility not only to get access to the informational resources but they can also take part in forming their own net content, thus we can see informational filling of internet;
- Mastering of informational conceptions, knowledge and habits. New possibilities open for activities where both lecturers and students can easily join even if they do not possess special knowledge on informatics;

- Observation over activity of the participants of community;
- Creation of learning situations where it is possible to observe and learn phenomena inaccessible before.

One of the most expended services of WEB 2.0 are the blogs – web-sites the main content of them are the notes that are regularly added.

Today the enormous quantity of instruments of communications based on the informational technologies is worked out; many of them find **addition in the context of learning.

Communication is the base of learning. Participation (or conversation) of students is not allowed during traditional kinds of learning. Students must also have real possibilities for publication of knowledge.

For lecturers, published works of students is a possibility to make conclusions about learning activity of students. Such kind of publication for students is a material for further reflection and analysis that permits them to come back to their works and to re-comprehend them once more, enriching their own learning experience. Besides, such kind of publication permits to get feedback which helps students in the process of construction of knowledge (Ferdig and Trammell, 2006). Using blogs is very useful in this aspect because they give students the possibility to comprehend researched material and to imagine its understanding.

Students can use blogs during learning any discipline. Here they can place learning materials, tasks, questions for self-checking and other learning information. It is interesting for students when the information in the blogs constantly renovates. It can be both separate news on proper discipline and important declarations.

Blogs can serve not only as means of organisation of the process of learning and communication between lecturers and students, but their thoughts or their additional materials can be shared with the group.

Before the process of beginning of using the blog in the pedagogical activity it is necessary to understand how it works and what possibilities we will have using this resource.

The aims of the creation of learning blogs can vary. It is useful to look through the blogs that are correlated with definite practice of teaching in order to see how they are used in the learning process.

Using blogs in the process of learning has a number of peculiarities, that is why it is important for students to get acquainted with the conception of blogs and the aims of their creation, with how these blogs work and to show the examples of successful and unsuccessful blogs etc. It is useful to work out strict rules for learning from a blog that will determine frequency of placing of reports, their volume, and the quantity of hyper sending, necessity to stick to the topic of discussion. These rules are created together with the students and it is necessary to speak about definite prohibitions: if communication in the blogs is unofficial online communication, students sometimes students use in their blogs too informal language and besides they do not always stick to the rules of design of quotations and sending to the sources (Ferdig and Trammell, 2006).

Students' blogs must be available not only for the students of the same group or course but also for wider public. With this aim it is also possible to involve ex-

perts and all volunteers into the processes of reading and commenting the students' blogs. Communication with the experts makes received knowledge not only more significant but it is evidence of the fact that real people read students' blogs. More open character of blogs will force students to spend more time during the preparation of information and will force them to concern more critically to themselves and to all information they write.

The Student must remember that once the text was published on the internet, it could be read immediately, and that from this moment on communication was irreversible even if later the information would be edited or removed.

Possibility of placing comments in the blogs favors to get feedback and potential support of new ideas and possibility of inculcation of hyper sending to the text and to other resources that helps students to comprehend interaction and context of knowledge, its construction and mastering.

When using blogs in the course of learning, it is necessary to include significant demands to the process of learning (to the addition to the tests). It is possible to get from the blog approximately one-third of all information about the process of learning.

For taking an active part in the learning process every student must write between five to nine information, of no less than 250 words each, during one semester. Every informational blog must have written skills he/she got in the classroom or in the process of self-dependent work.

In reality, most students write more information than necessary. They also read information of other people, comment on them and lecturer must do it, too. The process of learning becomes, at least, two-polar. First of all, students comprehend everything they learned and the facts that they 'silently' understood, place it in the form of a document on the blog. In short, they give examples and mistakes (or positive experience) that was researched by them during the process of learning. Secondly, by reading and commenting on other blogs, students begin to study side by side without participation of the lecturer, they try to get knowledge that they can share in the future.

Students that are in the blogosphere (net of blogs) are not limited in their access to systematised information and exchange of experience. Using blogs in pedagogical practice breaks easily all psychological barriers of communication and teaches students to communicate effectively.

Conclusion

Thus, with the aim of rise of effectiveness of learning at higher educational establishments it is necessary to use services WEB 2.0 more concretely. More and more, providing the most simple learning program is being replaced by the mechanism of creation of content where the process of learning itself is bearing; instead of the process of reading of learning materials that were prepared by the creators of courses beforehand, students and lecturers could create them themselves. It raises interest in learning and motivation, stabilises purposefulness and in the end improves higher education as a whole.

References

Блог як школа: Створення ефективного навчального блога! Частина 2. [Creation of an effective blog. Part 2] [on-line]. (2009): Українська блогосфера [Ukrainian blogosphere]. [cit. 2010]. Available at: http://blogosphere.com.ua/2009/03/02/how-to-create-effective-educational-blog/.

Decree of President of Ukraine 347/2002 from 17.04.02 about National doctrine of the development of education [on-line]. (2002): [cit. 2010]. Available at: http://osvita.ua/legislation/other/2827.

Ferdig, R. E., Trammell, K. D. (2006): Обучение в "блогосфере" [Content Delivery in 'blogosphere'] [on-line]. [cit. 2010]. Available at: http://www.itlt.edu.nstu.ru/article21_richard_ferdig_kaye_trammell.php.

Kuharenko, V. M., Rybalko O. V., Sirotinko N. G. (2002): *Distance learning: conditions of using*. [Distance course: Textbook]. 3rd ed. Kharkov : NTU 'KPI' 'Torsing', 320 pp.

O'Reilly, Tim (2005): What Is Web 2.0 [on-line]. [cit. 2005-09-30]. Available at: http://oreilly.com/web2/archive/what-is-web-20.html.
ibakceO'Reilly, Tim (2005): What Is Web 2.0 [on-line].. *Computerra*, 37 (609) and 38 (610) [cit. 2010].

O'Reilly, Tim (2005): Computerra online [on-line]. [cit. 2010]. Available at: http://www.computerra.ru/think/234100/.

Social Network as a Part of the Interactive Environment for Starting Entrepreneurs

David Sousedík
JVM-RPIC, s. r. o., Štefánikova 167, 760 30 Zlín, Czech Republic
e-mail: sousedik@jvmrpic.cz

Ladislav Buřita
Department of CIS, Faculty of Military Technology, University of Defence, Kounicova 65, 612 00 Brno, Czech Republic
Department of IEIS, Faculty of Management and Economics, Tomas Bata University, Mostní 5139, 760 01 Zlín, Czech Republic
e-mail: ladislav.burita@unob.cz

Abstract

The paper presents a concept of the interactive environment for starting entrepreneurs from the perspective of three defined areas: education, communication and cooperation. It is based on the outcomes of an ongoing research which surveys the opinions of prospective entrepreneurs on the electronic environment. The result is the processing and incorporation of the needs of those interested in the business interactive environment and the identification of areas that will be developed within this environment. A social network will be created as a part of the environment.

Key words

Interactive environment; social network; research; education; communication; cooperation; starting entrepreneurs

Project on interactive environment for starting entrepreneurs

Interactive environment for starting entrepreneurs is a project that is currently under development and preparation. The basic idea is to develop an interactive environment for those interested in business, which will facilitate their training, interaction and participation in its content at ipodnikatel.cz portal, which will house the environment. The interactive environment will provide starting entrepreneurs with a network which will include:

- Education(participation in an on-line course for starting a business),
- Communication (with those who are preparing themselves for starting a business),
- Cooperation (expressing opinions on the content of the www.ipodnikatel.cz portal).

All three areas have to be interconnected and create a synergy effect. Those interested participate in an on-line course which trains them for starting a business. At the same time they can interact with the community of people with the

same interest. They exchange their views and experiences which contributes to the educational effect. Ipodnikatel.cz portal covers the entire environment. Those interested in business must be a part of it and the content of the portal has to be tailored to their needs. The goal is to convert those interested in business into successful entrepreneurs [4].

Posing basic research questions

The interactive environment for starting entrepreneurs has to be based on real needs and ideas of its potential users. Therefore, it is necessary to engage them into the preparation stage of the project as soon as possible. (Sousedík, 2009) Currently, there is an ongoing research that is presented at http://www.hwsystem.ic.cz/, which is, in cooperation with `ipodnikatel.cz`, collecting views of those interested in business on the intended interactive environment for beginning entrepreneurs. The aim is to select appropriate Web 2.0 services (Zbiejczuk, 2007) which the prospective entrepreneurs would appreciate, and thus respond to the basic research questions raised prior to the research process:

- Will the prospective entrepreneurs be attracted by all three areas (education, communication and cooperation)?
- Will they at least take an interest in the on-line course?
- Will it be necessary to create their own social network for communication, or will they give priority to communication through the already existing social networks (e.g. Facebook)?
- What factor would motivate the prospective entrepreneurs to participate in the creation of the environment and regularly visit the portal which hosts this environment?

Ongoing research results

The questionnaire survey was commenced in December 2009 and is still ongoing. After the first three months the main trends resulting from the research can be already presented. The complete questionnaire is available at `www.hwsystem.ic.cz`.

Identifying questions and questions focused on entrepreneurship

In the part of identifying questions and questions focused on business the most numerous responses are presented only. This paper does not aspire to present a detailed analysis of the survey that is currently still ongoing. A total of 52 questionnaires were completed; 35 were answered by women and 17 by men. The largest groups of respondents were 26–30 and 36–40 years old. The highest education achieved by most respondents (38) was secondary vocational school (with the school leaving examination). Forty respondents are currently employed full time, 28 prospective entrepreneurs plan to start a business within one year. Most respondents (34) have no experience of entrepreneurship from the past.

Interactive environment

The third part of the questionnaire was focused on the intended interactive environment for starting entrepreneurs. The objective was to find out if the respondents were interested in the interactive environment services as a whole, or if they preferred just one of the areas. The results so far show that the most attractive area is education, namely the chance to participate in on-line courses for starting entrepreneurs, which will provide them with the knowledge and skills useful for starting a business. This was the choice of 23 respondents. 17 respondents answered that they would appreciate all offered areas of services as a whole (education, communication, cooperation). The area of communication was placed third (10 respondents) and only two prospective entrepreneurs found the separated area of cooperation the most attractive; in other words to have a chance to participate in the content of the `ipodnikatel.cz` portal, which houses the environment.

The results show that the interactive environment has to be built mainly on the area of education; at the same time it has to offer other services as well and function as an integral whole of the proposed areas. Although the cooperation area seems uninteresting, and probably would not work on its own, it remains – in connection with education and communication areas – an integral part of the whole. Preliminary results already answer one of the research questions. Within the environment, the areas of education, communication and cooperation can be developed together. Those interested in entrepreneurship have accepted all three areas, which should work as a whole, while the main emphasis has to be put on the education area.

Another part of the questionnaire deals with the development of the area of education. This offers the opportunity to engage starting entrepreneurs in an on-line course. The main success or failure of the whole environment will depend on the quality and practical implementation of this area. A total of 30 respondents said they would like to sign up for the on-line course; however, they would consider the course fee. The question elaborates on the structure of the on-line course for starting entrepreneurs. The areas, the respondents could take a stand on, were as follows:

- Through a series of questions (entrance test) to determine the areas where I need to improve.
- Participation in an e-learning course tailored to individual needs (development of the areas which I need to improve).
- Participation in an e-learning course aimed at gaining a general overview.
- On-line consultation with a personal adviser on a business prospectus.
- Business plan development with the assistance of a personal adviser (on-line communication).
- Creating a personal profile to be able to communicate with the others (e.g. other course participants).
- Involvement in group discussions moderated by an experienced consultant.
- Participation in the final test and obtaining a certificate of course completion.
- Subscription to a journal which will be sent by e-mail.

The most attractive areas of interest are:

1. On-line consultation with a personal adviser on the issue of the business prospectus. (28)
2. Through a series of questions to determine the areas in which I need to improve. (27)
3. Participation in an e-learning course tailored to individual needs. (27)
4. Participation in the final test and obtaining a certificate of course completion. (21)
5. Business plan development with the assistance of a personal adviser. (20)

The answers to this question confirm that the structure of the education area has been designed rightfully. Those interested in business understand this area as a whole consisting of an entrance test, subsequent training through an e-learning course, preparation of a business plan; they also require feedback in the form of a final test.

The attractiveness of working on the business plan on-line with the chance to consult it was confirmed by the question in which the respondents were allowed to select one answer only. The options 'On-line consultation on a business plan with a personal adviser' and 'Development of a business plan with a personal adviser (on-line communication)' selected by most respondents (13 and 11).

Thus the evaluation of the research questions is obvious. The opportunity to participate in an on-line course for starting entrepreneurs caught the respondent's interest, whereas the most attractive option was the on-line business plan development with a personal adviser.

Communication

Concerning the first question, most respondents (43) chose the option 'Communication within the interactive environment'. In this way they answered another research question. When creating an interactive environment, it is necessary to develop one's own social network. The respondents did not show any interest in any other social networks (e.g. Facebook). This is a fairly surprising result which has shifted the area of communication to a new level. The chance to communicate does not play a complementary role only, but it is a pivotal pillar of the interactive environment for starting entrepreneurs.

Those interested in business would create their personal profile (29 responses) and most often they would like to communicate with other people interested in business (9) and experts in a particular area of business (law, marketing, accounting, etc.) (6). The area of communication should also include additional services that add value to it. Apart from the ability to communicate those interested in business would also appreciate the following services:

- Creating a presentation about their business activities and its confrontation within the social network (26).
- Participation in on-line discussions with experts (business, accounting, etc.) (26)3.
- Engagement in real time discussions (chat) (25).

Several respondents, who were not interested in creating their personal profile, expressed their interest in communication through chat. It opens up another issue which will be necessary to be dealt with when creating the interactive environment.

Cooperation

Another part of the survey identifies the interest in the opportunity to be involved in the creation of the `www.ipodnikatel.cz` portal which hosts the entire interactive environment. A total of 21 respondents expressed their interest in cooperation. Among the most common forms of cooperation were scoring or grading individual papers presented at the portal (12) and selecting the best contributions of the month / week / day (9). However, the greatest number of responses in this section was the answer 'I do not know' (24) by which the interest in co-operation was neither confirmed nor invalidated.

Within the area of cooperation there was also a question on the form of motivation which would attract the prospective entrepreneurs to visit the `www.ipodnikatel.cz` site regularly. The results show that the major motivation would be the chance to contact the experts and ask questions on the issues related to starting a business. This option has been chosen by 33 respondents (out of 52). It answers the last research question seeking to find the most motivating factor that would attract the prospective entrepreneurs into the participation in the environment development.

Main results and identification of further research areas

The interactive environment for starting entrepreneurs is a combination of three basic areas, namely education, communication and cooperation. This environment is accessible on-line which significantly reduces the cost of preparing the starting entrepreneurs. Otherwise it is very expensive and requires public support. The aim of the interactive environment concept is to support people interested in becoming successful starting entrepreneurs while maintaining the profitability of this project. The outcomes of the ongoing research have showed the following findings:

- The respondents expressed their interest in all three areas that should function as an integral whole. The main emphasis has to be put on the area of education.
- The opportunity of participating in an on-line course for starting entrepreneurs caught the respondent's interest; the most attractive option was the on-line business plan development with a personal adviser.
- When creating the interactive environment it is necessary to build a social network. The respondents were not interested in any other social network (e.g. Facebook).
- The opportunity to contact and consult experts about issues related to the start of business is the main motivating factor that would attract the participants to use the portal housing the interactive environment, and thus make them interested in its development.

Within future research it is necessary to elaborate on further individual areas of the interactive environment:

- Methodology, content and form of implementation of the on-line course for starting entrepreneurs.
- Designing and creating a social network for interactive environment using technologies by Buřita and Janoušek (2007).
- Forms of cooperation within the portal housing the environment.

References

Sousedík, David (2009): Jak využít služeb Webu 2.0 při tvorbě interaktivního prostředí pro začínající podnikatele. In: *II. mezinárodní vědecká konference doktorandů a mladých vědeckých pracovníků, díl II.* Karviná (CZ) : Slezská univerzita, pp. 1083–1087. (ISBN 978-80-7248-553-6.)

Buřita, Ladislav; Janoušek, Michal (2007): Semantic Web Ontology and Technology. In: *CD Proceedings of the Military CIS Conference – MCC 2007.* Bonn : FGAN FKIE, pp. 55–58. (ISBN 978-3-934401-16-7.)

Zbiejczuk, Adam (2007): *WEB 2.0 – charakteristiky a služby.* (Diploma thesis.) Brno : Masarykova univerzita 71 pp. Available at: `http://www.zbiejczuk.com/web20/`.

Stephenson, James (2008): 25 Common Characteristics of Successful Entrepreneurs [on-line]. [cit. 2010]. Available at: `http://www.entrepreneur.com/homebasedbiz/article200730.html`.

The contribution is a part of the ongoing thesis 'Knowledge Support for Starting Entrepreneurs through Information and Communication Technologies' at Tomas Bata University in Zlin by David Sousedík. It also draws from research topics on the Knowledge Systems dealt within the MENTAL project at the University of Defence in Brno, which is being solved by the co-author of this article, Ladislav Buřita.

Call for Papers

Papers to be included in the next issue should be preferably focused on topics related to social-networks in one or more of the following subjects (the list is indicative rather than exhaustive):

Sentiment/Opinion Analysis in Natural-Language Text Documents

Algorithms, Methods, and Technologies for Building and Analysing Social Networks

Applications in the Area of Social Activities

Knowledge Mining and Discovery in Natural Languages Used in Social Networks

Medical, Economic, and Environmental Applications in Social Networks

Submitted papers should not have been previously published nor be currently under consideration for publication elsewhere. Each of the submitted research papers should not exceed 26 pages. All papers are refereed through a peer review process.

Submissions should be send in the PDF form via email to the following address: `SoNet.RC@gmail.com`

Accepted papers are to be prepared according to the instructions available at `http://www.konvoj.cz/journals/mmm/`.